

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平8-77294

(43) 公開日 平成8年(1996)3月22日

(51) Int.Cl.⁸

G 0 6 K 9/46

9/20

9/62

識別記号

G

3 4 0 L

6 1 0 Z

庁内整理番号

9061-5H

9061-5H

F I

技術表示箇所

審査請求 未請求 請求項の数10 O L (全 30 頁)

(21) 出願番号 特願平6-212951

(22) 出願日 平成6年(1994)9月6日

(71) 出願人 000003078

株式会社東芝

神奈川県川崎市幸区堀川町72番地

(72) 発明者 石谷 康人

東京都青梅市末広町2丁目9番地 株式会

社東芝青梅工場内

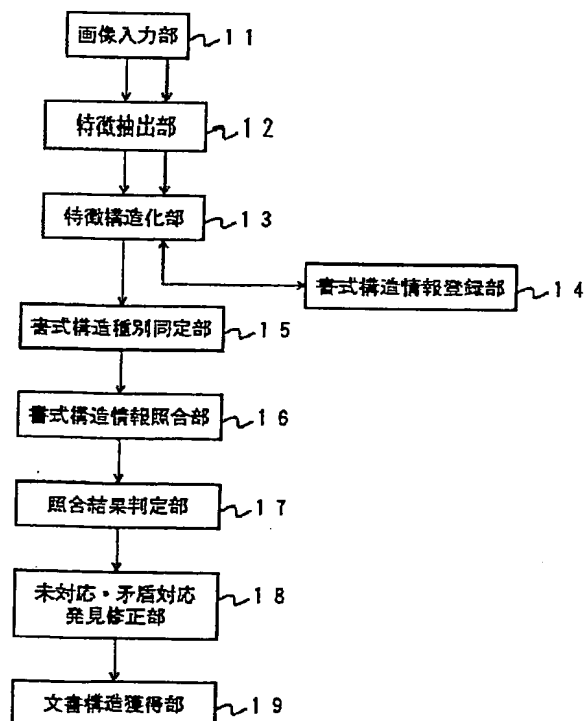
(74) 代理人 弁理士 鈴江 武彦

(54) 【発明の名称】 文書画像処理装置

(57) 【要約】

【目的】 本発明は、帳票などの文書フォーマットを正確に特定でき、効率の良い文字列の抽出、読み取りを可能にした文書画像処理装置を提供する。

【構成】 画像入力部 11 より生成される帳票の入力画像に対して、特徴抽出部 12 より抽出された図形特徴量を特徴構造化部でグループ化し、それぞれの特徴間の関係を抽出・管理する。構造化特徴と、書式構造種別同定部 15 で予め登録されている処理対象文書の書式構造に関する情報（書式構造モデル）を用いて入力文書の書式構造の種別を推定する。書式構造情報照合部 16 は、推定された書式構造の種別に対応する書式構造モデルと入力文書の構造化特徴の間で、詳細な対応関係を抽出する。未対応・矛盾対応発見修正部 18 で対応関係の整合を得た後、文書構造獲得部 19 でその対応関係に基づき予め登録されている書式構造モデルに関する情報を入力文書にコピーすることで入力文書の構造及び関連知識を獲得する。



【特許請求の範囲】

【請求項 1】 文書より入力画像を生成する画像入力手段と、

入力画像の書式構造を認識するために用いられる処理対象文書の書式構造に関する情報（書式構造モデル）を予め登録する書式構造情報登録手段と、

前記画像入力手段により生成された入力画像から幾何学的な図形特徴量を抽出する特徴抽出手段と、

前記特徴抽出手段より抽出された図形特徴量をグループ化して画像特徴を生成し、それぞれの画像特徴間の関係を抽出・管理する特徴構造化手段と、

前記特徴構造化手段で得られた入力画像の画像特徴と、

前記書式構造情報登録手段によって予め登録されている処理対象文書の書式構造に関する情報を用いて、入力文書の書式構造の種別の候補を絞りこむ書式構造種別同定手段と、

前記書式構造同定手段で候補となったすべての書式構造モデルと前記特徴構造化手段で構造化された入力文書の特徴との間で対応付けを行ない、最も良く対応づいた書式構造モデルと入力文書の組を選択し、その対応関係を獲得する書式構造情報照合手段と、

前記書式構造情報照合手段で選択された書式構造文書と入力文書における構造化特徴間の対応付けにおいて、不完全な対応付けおよび矛盾した対応付けを解消することにより整合のとれた前記書式構造モデルと入力文書の構造化された特徴間の対応関係を獲得する未対応・矛盾対応発見修正手段と、

前記未対応・矛盾対応発見修正手段によって得られた前記書式構造モデルと入力文書の構造化された特徴間の対応関係に基づいて、予め登録されている当該書式構造モデルに関する情報を入力文書にコピーすることにより入力文書の書式構造と関連情報を獲得する文書構造獲得手段と、

を具備することを特徴とする文書画像処理装置。

【請求項 2】 文書より入力画像を生成する画像入力手段と、

前記画像入力手段により生成された入力画像から線分と文字成分に関する図形特徴を抽出し、さらに前記入力画像における文字成分以外の領域から線分に関する特徴を罫線を構成する図形特徴とみなして抽出する特徴抽出手段と、

前記特徴抽出手段より抽出された罫線に関する図形特徴をグループ化することにより表に関する特徴を抽出し、各表に関する特徴において罫線が交差・接続する部分に生じる接合部に関する情報を抽出し、それぞれの特徴間の関係を抽出・管理する特徴構造化手段と、

を具備することを特徴とする文書画像処理装置。

【請求項 3】 文書より入力画像を生成する画像入力手段と、

前記画像入力手段により生成された入力画像から線分と

文字成分に関する図形特徴を抽出し、さらに前記入力画像における文字成分以外の領域から線分に関する特徴を罫線を構成する図形特徴とみなして抽出する特徴抽出手段と、

前記罫線特徴抽出手段より抽出された罫線に関する図形特徴の集合に対して、交差・接続する罫線を同じグループにまとめることにより表に関する特徴を抽出する表特徴抽出手段と、

を具備することを特徴とする文書画像処理装置。

【請求項 4】 文書より入力画像を生成する画像入力手段と、

入力画像の書式構造を認識するために用いられる処理対象文書の書式構造に関する情報（書式構造モデル）を予め登録する書式構造情報登録手段と、

前記画像入力手段によって生成された入力画像から線分と文字成分に関する図形特徴を抽出し、さらに入力画像における文字成分以外の領域から線分に関する特徴を罫線に関する図形特徴とみなして抽出する特徴抽出手段と、

前記特徴抽出手段より抽出された罫線に関する図形特徴をグループ化することにより表に関する特徴を抽出し、各表に関する特徴において罫線が交差・接続する部分に生じる接合部に関する情報を抽出し、それぞれの特徴間の関係を抽出・管理する特徴構造化手段と、

前記特徴構造化手段により得られた前記入力文書の表に関する特徴と、予め前記書式構造情報登録手段により登録されている書式構造モデルを構成する表に関する特徴との間で照合処理を行い、表間対応関係を獲得する表照合手段と、

前記表照合手段により得られた表の対応関係において入力文書の表を構成する罫線と、同罫線に対応付く書式構造モデルの表を構成する罫線との間の対応関係を獲得する罫線照合手段と、

前記照合処理結果に基づき特徴間の対応付きの程度を表す照合度を計算し、正しい対応付けが行なわれているか否かの判断を行なう照合結果判定手段と、

を具備することを特徴とする文書画像処理装置。

【請求項 5】 文書より入力画像を生成する画像入力手段と、

入力画像の書式構造を認識するために用いられる処理対象文書の書式構造に関する情報（書式構造モデル）を予め登録する書式構造情報登録手段と、

前記画像入力手段によって生成された入力画像から幾何学的な図形特徴量を抽出する特徴抽出手段と、

前記特徴抽出手段より抽出された図形特徴量をグループ化して画像特徴を生成し、それぞれの特徴間の関係を抽出・管理する特徴構造化手段と、

前記特徴構造化手段で得られた入力画像の構造化された画像特徴と、予め前記書式構造情報登録手段によって登録されている処理対象文書の書式構造に関する情報を用

いて、類似度を計算し、最も類似度の高い書式構造モデルあるいは類似度の高いものから順に複数個の書式構造モデルあるいはある一定値以上の類似度を有する書式構造モデルを選び、前記入力文書の書式構造の種別を一つあるいは複数個の候補に絞りこむ書式構造種別同定手段と、

を具備することを特徴とする文書画像処理装置。

【請求項 6】 文書より入力画像を生成する画像入力手段と、

入力画像の書式構造を認識するために用いられる処理対象文書の書式構造に関する情報（書式構造モデル）を予め登録する書式構造情報登録手段と、

前記画像入力手段により得られた入力画像から線分と文字成分に関する図形特徴を抽出し、さらに入力画像における文字成分以外の領域から線分に関する特徴を罫線に関する図形特徴とみなして抽出する特徴抽出手段と、

前記特徴抽出手段より抽出された罫線に関する図形特徴をグループ化することにより表に関する特徴を抽出し、各表に関する特徴において罫線が交差・接続する部分に生じる接合部に関する情報を抽出し、それぞれの特徴間の関係を抽出・管理する特徴構造化手段と、

前記特徴構造化手段で得られた入力画像の構造化された特徴と、予め前記書式構造情報登録手段によって登録されている処理対象文書の書式構造に関する情報を用いて、類似度を計算し、最も類似度の高い書式構造モデルあるいは類似度の高いものから順に複数個の書式構造モデルあるいはある一定値以上の類似度を有する書式構造モデルを選び、入力文書の書式構造の種別を一つあるいは複数個の候補に絞りこむ書式構造種別同定手段と、前記書式構造種別同定手段により選択されたそれぞれの書式構造モデルに対して、

前記特徴構造化手段により得られた入力文書の表に関する特徴と、前記書式構造情報登録手段により登録されている当該書式構造モデルを構成する表に関する特徴との間照合処理を行ない、表間対応関係を獲得する表照合手段と、

前記表照合手段により得られた表の対応関係において入力文書の表を構成する罫線と同罫線に対応付く書式構造モデルの表を構成する罫線との間の対応関係を獲得する罫線照合手段と、

前記罫線照合手段により得られた対応関係に対して、特徴間の対応付きの程度を表す照合度を計算する照合度計算手段と、

前記照合度計算手段で計算されたそれぞれの書式構造モデルの照合度の中から最大照合度を示す書式構造モデルを抽出する照合結果出力手段と、

前記照合結果出力手段によって抽出された書式構造モデルの最大照合度を用いて、入力文書と当該書式構造モデルの構造化特徴間で正しい対応付けが行われているか否かの判定を行なう照合結果判定手段と、

を具備することを特徴とする文書画像処理装置。

【請求項 7】 前記書式構造情報登録手段は、入力文書の書式構造を認識するために用いられる処理対象文書の書式構造に関する情報を登録する際に、正立した処理対象文書の書式構造に関する情報を複数の所定角度で回転させたものを発生させ、それぞれに正立したものから何度回転しているかに関する情報を付与し、それらすべてを処理対象文書の書式構造に関する情報として登録することを特徴とする請求項 1 または請求項 4 または請求項 5 または請求項 6 記載の文書画像処理装置。

【請求項 8】 文書より入力画像を生成する画像入力手段と、

入力文書の書式構造を認識するために用いられる処理対象文書の書式構造に関する情報を登録する際に、正立した処理対象文書の書式構造に関する情報を複数の所定角度で回転させたものを発生させ、それぞれに正立したものから何度回転しているかに関する情報を付与し、それらすべてを処理対象文書の書式構造に関する情報として登録する書式構造情報登録手段と、

前記画像入力手段により生成された入力画像から幾何学的な図形特徴量を抽出する特徴抽出手段と、

前記特徴抽出手段より抽出された図形特徴量をグループ化することで画像特徴を生成し、それぞれの特徴間の関係を抽出・管理する特徴構造化手段と、

前記特徴構造化手段で得られた入力画像の構造化された画像特徴と、前記書式構造情報登録手段によって予め登録されている処理対象文書の書式構造に関する情報を用いて、入力文書の書式構造の種別を一つあるいは複数個の候補に絞りこむ書式構造種別同定手段と、

前記書式構造種別同定手段で候補となったすべての書式構造モデルと入力文書の前記特徴構造化手段で構造化された特徴との間に対応付けを行ない、最も良く対応づいた書式構造モデルと入力文書の組を選択し、その対応関係を獲得するモデル照合手段と、

前記モデル照合手段で選択された書式構造文書と入力文書における構造化特徴間の対応付けにおいて、不完全な対応付けおよび矛盾した対応付けがなされているか否かを発見し、それらを解消することにより整合のとれた前記書式構造モデルと入力文書の構造化された特徴間の対応関係を獲得する未対応・矛盾対応発見・修正手段と、

前記モデル照合手段で入力画像に対応付いた書式構造モデルが所定角度で回転させたものである場合には、その回転角度を正立する方向に入力画像を回転し、正立した当該書式モデルと対応付ける画像回転手段と、

前記書式構造モデルと入力文書の構造化された特徴間の対応関係に基づいて、予め登録されている当該書式構造モデルに関する情報を入力文書にコピーすることにより入力文書の書式構造と関連情報を獲得する文書構造獲得手段と、

を具備することを特徴とする文書画像処理装置。

【請求項 9】 文書より入力画像を生成する画像入力手段と、

入力画像の書式構造を認識するために用いられる処理対象文書の書式構造に関する情報（書式構造モデル）を予め登録する書式構造情報登録手段と、

前記画像入力手段によって生成された入力画像から線分と文字成分に関する図形特徴を抽出し、さらに入力画像における文字成分以外の領域から線分に関する特徴を罫線に関する図形特徴（罫線特徴）とみなして抽出する特徴抽出手段と、

前記特徴抽出手段より抽出された罫線特徴をグループ化することにより表に関する特徴（表特徴）を抽出し、各表特徴において罫線が交差・接続する部分に生じる接合部に関する情報を抽出し、それぞれの特徴間の関係を抽出・管理する特徴構造化手段と、

前記特徴構造化手段により得られた入力文書の表特徴と、予め前記書式構造情報登録手段により登録されている書式構造モデルを構成する表特徴との間で照合処理を行い、表間対応関係を獲得する表照合手段と、

前記表照合手段により得られた表の対応関係において入力文書の表を構成する罫線特徴とそれに対応付く書式構造モデルの表を構成する罫線との間の対応関係を獲得する罫線照合手段と、

前記罫線照合手段により獲得された対応関係に対し、特徴間の対応付きの程度を表す照合度を計算し、正しい対応付けが行なわれているか否かの判定を行なう照合結果判定手段と、

前記照合結果判定手段によって正して対応付けが行われていると判定された入力文書と書式構造モデルの罫線特徴間の対応関係において、入力文書の罫線特徴（入力罫線特徴）に対応付いていない書式構造モデルの罫線特徴（未対応モデル罫線特徴）を抽出する手段と、

前記未対応モデル罫線特徴に対応付くべき入力罫線特徴が他のモデル罫線特徴に対応付いている場合には、その対応関係を解消し、未対応モデル罫線特徴とその入力罫線特徴を対応付ける手段と、

前記未対応モデル罫線特徴に対応付くべき欠損した未対応の入力罫線特徴がある場合には、未対応モデル罫線特徴と当該入力罫線特徴を対応付ける手段と、

前記未対応モデル罫線特徴に対応付くべき入力罫線特徴が見つからない場合には、対応付くべき入力罫線特徴を新たに発生させ、未対応モデル罫線特徴と新たに発生させた入力罫線特徴を対応付ける手段によって未対応および矛盾対応を修正する手段と、

を具備することを特徴とする文書画像処理装置。

【請求項 10】 文書より入力画像を生成する画像入力手段と、

入力画像の書式構造を認識するために用いられる処理対象文書の書式構造に関する情報（書式構造モデル）を予め登録する書式構造情報登録手段と、

前記画像入力手段により生成された入力画像から線分と文字成分に関する図形特徴を抽出し、さらに入力画像における文字成分以外の領域から線分に関する特徴を罫線に関する図形特徴（罫線特徴）とみなして抽出する特徴抽出手段と、

前記特徴抽出手段より抽出された罫線特徴をグループ化することにより表に関する特徴（表特徴）を抽出し、各表特徴において罫線が交差・接続する部分に生じる接合部に関する情報を抽出し、それぞれの特徴間の関係を抽出・管理する特徴構造化手段と、

前記特徴構造化手段により得られた入力文書の表特徴と、予め書式構造情報登録手段により登録されている書式構造モデルを構成する表特徴との間で照合処理を行い、表間対応関係を獲得する表照合手段と、

前記表照合手段により得られた表の対応関係において入力文書の表を構成する罫線特徴とそれに対応付く書式構造モデルの表を構成する罫線特徴との間の対応関係を獲得する罫線照合手段と、

前記罫線照合手段で獲得された罫線特徴間の対応関係において、当該書式構造モデルの罫線特徴集合における罫線間の接続関係に基づいて、対応する入力文書の罫線特徴集合における罫線間の接続関係を修正する罫線照合後処理手段と、

を具備することを特徴とする文書画像処理装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、例えば帳票などの文書上に記載された文字の読みとり、データベースへの自動入力、帳票画像の自動ファイリングに用いられる文書画像処理装置に関するものである。

【0002】

【従来の技術】近年、書類形態で管理・利用されていた文書を電子化して計算機に入力し、多様な用途で活用している。様々な業務処理で用いられている文書には、表形式のものが多く、文書中の特定の位置に記載されている所望の文字列を所望の形式で計算機に効率よく入力し、管理したいという要求が高まっている。これらの期待に答えるためにこれまでに光学的文字読みとり装置が開発され、実用化されてきた。

【0003】このような帳票上の特定の位置に記載されている文字列を読み取り対象とする場合、帳票の書式構造を事前に知る必要がある。帳票のように定型的な書式構造を持つ文書に対する文字列の読み取り処理では、書式構造を事前に覚えさせておき、それに基づいて効率よく文字列領域を特定し、読み取るようにしている。

【0004】このようなアプローチをとるものとしては、対象文書の書式構造に関する知識とそれを利用する処理を分離することで書式構造の多用性への拡張性を高めている。これらのものは書式構造に関する知識として文書構造の物理情報、例えば位置、大きさ、幾何学的関

係などの情報を用いている。代表的な研究として、文献「信学論(D)、J71-D、10、pp.2050-2058(1988-10)」と文献「信学論(D-II)、J72-D-II、7、p.1029-1039(1989-07)」がある。

【0005】一方、最近になって、文書構造に関する物理情報を用いるだけでは対応できない帳票に対しても文字列領域の特定を可能にすることが考えられている。この場合、文字領域を空間的な隣接・接続関係に基づいた論理情報で表現し、識別することを可能にしている。

【0006】このようなアプローチをとるものとしては、種々の帳票文書の書式構造に関する構成規則を一般化してメタ知識と表現されたものを用いて対象文書画像の書式構造を認識している。代表的な研究として文献「信学論(D-II)、J76-D-II、3、pp.534-545(1993-03)」がある。

【0007】

【発明が解決しようとする課題】文献「信学論(D)、J71-D、10、pp.2050-2058(1988-10)」や文献「信学論(D-II)、J72-D-II、7、pp.1029-1039(1989-07)」のようなアプローチをとる手法では、物理情報に依存しているので文書が大幅にずれて入力されたり、拡大・縮小によるスケール変換を受けている場合には対応できない。また、文献「信学論(D-II)、J76-D-II、3、pp.534-545(1993-03)」のようなアプローチをとる手法では、入力文書は高品質で画像の劣化がないことを前提としているため、処理対象文書の品質が悪く画像情報が不足している場合には所望の処理結果を得られないという問題がある。

【0008】また、何れの手法においても、以下の

(1)～(6)の問題がある。

【0009】(1) 表が複数混在している場合には対応できない。

【0010】(2) 表が分裂している場合には対応できない。

【0011】(3) 罫線がかすれていたり、欠落している場合には対応できない。

【0012】(4) 罫線分布が局所的に変動している場合には対応できない。

【0013】(5) 入力された文書の処理方法を限定している(左右に90度および180度回転している文書に対応できない)。

【0014】(6) 種々の書式構造の文書を一括して読み取らせる場合、適用すべき書式構造モデルの自動同定ができない。

【0015】本発明は上記事情に鑑みてなされたもので、表形式の帳票などの書式構造を正確に認識でき、効率の良い文字列の領域の特定を可能にした文書画像処理装置を提供することを目的とする。

【0016】

【課題を解決するための手段】本発明は、文書より入力

画像を生成する画像入力手段と、入力文書の書式構造を認識するために用いられる処理対象文書の書式構造に関する情報を登録する際に、正立した処理対象文書の書式構造に関する情報を複数の所定角度で回転させたものを発生させ、それぞれに正立したものから何度回転しているかに関する情報を付与し、それらすべてを処理対象文書の書式構造に関する情報として登録する書式構造情報登録手段と、前記画像入力手段により生成された入力画像から線分と文字成分に関する図形特徴を抽出し、さらに前記入力画像における文字成分以外の領域から線分に関する特徴を罫線を構成する図形特徴とみなして抽出する特徴抽出手段と、前記特徴抽出手段より抽出された罫線に関する図形特徴をグループ化することにより表に関する特徴を抽出し、各表に関する特徴において罫線が交差・接続する部分に生じる接合部に関する情報を抽出し、それぞれの特徴間の関係を抽出・管理する特徴構造化手段と、前記特徴構造化手段で得られた入力画像の構造化された画像特徴と、予め前記書式構造情報登録手段によって登録されている処理対象文書の書式構造に関する情報を用いて、類似度を計算し、最も類似度の高い書式構造モデルあるいは類似度の高いものから順に複数の書式構造モデルあるいはある一定値以上の類似度を有する書式構造モデルを選び、前記入力文書の書式構造の種別を一つあるいは複数の候補に絞りこむ書式構造種別同定手段と、前記特徴構造化手段により得られた入力文書の表に関する特徴と、前記書式構造情報登録手段により登録されている当該書式構造モデルを構成する表に関する特徴との間照合処理を行ない、表間対応関係を獲得する表照合手段と、前記表照合手段により得られた表の対応関係において入力文書の表を構成する罫線と罫線に対応付く書式構造モデルの表を構成する罫線との間の対応関係を獲得する罫線照合手段と、前記照合処理結果に基づき特徴間の対応付きの程度を表す照合度を計算し、正しい対応付けが行なわれているか否かの判断を行なう照合結果判定手段とにより構成されているモデル照合手段と、前記モデル照合手段で選択された書式構造文書と入力文書における構造化特徴間の対応付けにおいて、不完全な対応付けおよび矛盾した対応付けを解消することにより整合のとれた前記書式構造モデルと入力文書の構造化された特徴間の対応関係を獲得する未対応・矛盾対応発見修正手段と、前記モデル照合手段で入力画像に対応付いた書式構造モデルが所定角度で回転させたものである場合には、その回転角度を正立する方向に入力画像を回転し、正立した当該書式モデルと対応付ける画像回転手段と、前記書式構造モデルと入力文書の構造化された特徴間の対応関係に基づいて、予め登録されている当該書式構造モデルに関する情報を入力文書にコピーすることにより入力文書の書式構造と関連情報を獲得する文書構造獲得手段とを具備し、この文書構造獲得手段により得られる結果に基づいて入力画像の書式構造を

認識するように構成されている。

【0017】

【作用】この結果、本発明は、入力文書画像から罫線に関する図形特徴量と文字成分に関する図形特徴量を抽出し、これらのかかり合いから文字成分以外の領域で正確に罫線の特徴量を抽出できる。また、この罫線特徴をグループ化することにより表に関する特徴量が得られ、さらに各表において罫線に関する特徴とそれらの交差・接合により生じる接合部を抽出し、全体一部分関係で記述・管理することにより図形特徴量に対してより豊富な情報を付加することができる。それらを効率的に検索することができる。これにより文書内に表が複数個混在していても各々の情報を抽出することができ、また罫線に関する特徴も表ごとに区別することができる。

【0018】本発明では、入力画像から得られた構造化された特徴と予め登録されている書式構造モデルの構造化特徴との間で照合処理を行なうことにより入力文書を解釈するが、照合処理の前に書式構造モデルの種別を同定することで登録されているすべての書式構造モデルとの照合処理を行なうことを避けることができる。この処理では、文書中に含まれている接合部の数や罫線の数を用いているので入力文書と書式構造モデルの間で拡大縮小に伴うスケール変換がなされていたり、表や罫線の構成要素の大きさが部分的に変動していても安定した結果を得ることができる。書式構造モデルの種別を同定する時に唯一の結果を出力するのではなく複数の候補を出力することで同定誤りが生じないようにしている。候補となっている書式構造モデルの構造化特徴と入力文書の構造化特徴との間で行なわれる照合処理では、もっとも良く対応付く書式構造モデルを選ぶことができ、得られた対応関係が正しく獲得されているか否かの判断を行なうことにより処理結果の信頼性を高めている。このとき、照合処理まず表単位に行なわれ、次いで罫線単位に行なわれる。この結果、局所的にまったく同じ特徴量を持つ罫線でもそれが所属する表が異なれば誤った対応付けが行なわれることはない。

【0019】また、本発明では、対応関係に対してどちらかの構造化特徴の欠落による不完全な対応付きや矛盾した対応付きの有無およびこの箇所を発見し、それを解消する。この結果、整合のとれた対応関係を獲得することができ、安定した処理結果を得ることができる。この対応関係に基づいて予め登録されている書式構造モデルに関連する情報を入力文書に複写することにより入力文書の文書構造を獲得し、読みとるべき文字列の記載位置を正確に特定することができる。

【0020】

【実施例】以下、図面を参照して本発明の一実施例を説明する。図1は本実施例に係わる画像処理装置の概略構成を示すブロック図である。

【0021】本発明は、画像入力部11、特徴抽出部1

2、特徴構造化部13、書式構造情報登録部14、書式構造種別同定部15、書式構造情報照合部16、未対応・矛盾対応発見修正部17、照合結果判定部18、及び文書構造獲得部19の各機能部から構成されている。

【0022】また、処理対象とする帳票は、図2に示すように、罫線により文字列領域が規定されているものとする。

【0023】画像入力部11は、スキャナ装置やTVカメラ、FAXデータの入力部などによって構成されるものである。画像入力部11は、処理対象である帳票等の文書の画像を検出してシステム内に取り込む。画像入力部11によって入力された画像（入力画像）は、特徴抽出部12に送られる。

【0024】特徴抽出部12は、画像入力部11から得られた入力画像から、罫線や文字成分に関する幾何学的な図形特徴量をそれぞれ基本特徴として抽出する。特徴抽出部12により抽出された基本特徴の集合は特徴構造化部13に送られる。

【0025】特徴構造化部13は、特徴抽出部12により抽出された基本特徴の集合について、例えば罫線に関する特徴をグループ化することにより得られる表や、罫線が交差・接続する部分に生じる接合部などに関する情報を構造化特徴として抽出し、さらにそれぞれの特徴間の関係を記述、管理する。

【0026】書式構造情報登録部14は、システムとオペレータとの対話的な入力作業により、処理対象文書に関する知識の登録を行なう。本実施例では、処理対象文書の種類ごとに用意された種々の知識の総体をモデルと呼び、モデルを定義したときに用いた文書をモデル文書と呼ぶことにする。モデルは例えば、そのモデル文書の構造化特徴に関する情報などを有している。

【0027】このとき、書式構造情報登録部14は、正立している構造化特徴から、右に90度回転したもの、左に90度回転したもの、180度回転したものの3種類の構造化特徴を発生させ、それぞれが正立したものから何度回転されているものであるかという情報を付与して知識ベースとして格納してもよい。

【0028】この他に、例えば罫線により規定されている領域内に含まれる文字列を文字認識装置で認識・コード化させるような場合には、各種情報を構造化特徴に関連付けて知識ベースとして登録・管理するようにしてもよい。

【0029】各種情報には、例えば、文字認識対象領域の指定に関する情報、文字認識対象領域の文字列方向の情報、文字読み取り処理時に拘束条件として働く文字種情報及び筆記形態情報、文字認識後処理のための項目属性情報、文字認識結果の対応関係などの情報がある。

【0030】書式構造種別同定部15は、特徴構造化部13によって抽出された入力画像に対する構造化特徴と、予め書式構造情報登録部14によって登録されてい

るモデル文書のそれぞれの構造化特徴との間で、類似度計算を行なうことにより入力文書のフォーマットの種別を同定する。すなわち類似度値が高いモデルのフォーマットが入力文書のフォーマットである可能性が高いものと判別する。書式構造種別同定部 15 は、該当する構造化特徴を次段の書式構造情報照合部 16 に出力する。

【0031】なお、書式構造種別同定部 15 は、最も類似度値の高いモデルのみに注目するのではなく、例えばある事前に定められたしきい値以上の類似度値を示す全てのモデル文書の構造化特徴を候補として、書式構造情報照合部 16 に入力構造化特徴と共に送り込むようにしてもよい。

【0032】書式構造情報照合部 16 は、書式構造種別同定部 15 によって候補とされた全てのモデル文書と入力文書の間で構造化特徴間の対応関係を獲得する。また、書式構造情報照合部 16 は、入力文書の構造化特徴と各モデル文書の構造化特徴との対応付け結果に対して対応付きの度合いを表す尺度（以後、照合度と呼ぶ）を計算する。書式構造情報照合部 16 は、構造化特徴間の対応関係を示す情報、及び照合度を示す情報を照合結果判定部 17 に出力する。

【0033】なお、書式構造種別同定部 15 から複数のモデル文書の構造化特徴が送られてきた場合には、書式構造情報照合部 16 は、さらに最大照合度を示すモデル文書を選択・出力するようになっていてもよい。

【0034】照合結果判定部 17 は、書式構造情報照合部 16 によって得られた情報に基づいて、最大照合度を示す文書モデルと入力文書との間で対応関係が取れているか否かを判定する。ここで対応関係が取れていると判定された場合には、後段の未対応・矛盾対応発見修正部 18 に着目入力文書とモデル文書間の対応関係に関する情報が送られるようにする。対応関係が取れていないとみなされる場合には、入力文書を棄却して、次の文書を入力するようにオペレータに促す。

【0035】未対応・矛盾対応発見修正部 18 は、書式構造情報照合部 16 で獲得された入力文書の構造化特徴とモデル文書の構造化特徴の対応関係において、誤った特徴の抽出や必要な特徴の欠落のために不完全な対応や矛盾した対応が生じているか否かを発見する。未対応・矛盾対応発見修正部 18 は、不完全な対応や矛盾した対応を発見した場合には、それらを解消することにより整合のとれた入力・モデル間の対応関係を獲得した上で、文書構造獲得部 19 に出力する。

【0036】文書構造獲得部 19 は、未対応・矛盾対応発見修正部 18 で得られた整合のとれた入力文書とモデル文書の構造化特徴間の対応関係に基づいて、予め書式構造情報登録部 14 で登録されているモデル文書に関する知識を入力文書にコピーすることにより入力文書の文書構造及び関連知識を獲得する。

【0037】次に、上述した必須構成要素を含む具体的

な文書画像処理システムについて説明する。ここでは、図 3 に示すように、文字認識装置と組み合わせた文書画像処理システムについて説明する。このシステムは、入力文書から得られた入力画像において、特定の位置に記載されている文字列領域を自動的に抽出し、さらにその文字列画像を認識・コード化して、所望の出力様式で出力するという動作をするものである。以後、図 2 に示す文書（以下、特に断らない限り図 2 の文書を入力文書と呼ぶ）を用いて説明する。

【0038】図 3 に示すように、文書画像処理システムは、画像入力部 21、2 値化処理部 23、前処理部 25、特徴抽出部 27、特徴構造化部 29、モデル登録部 31、フォント種別同定部 33、モデル照合部 35、照合結果判定部 37、未対応矛盾対応発見修正部 39、文書構造獲得部 41、文字列領域抽出部 43、文字認識部 45、及び文字認識結果出力部 47 によって構成されている。

【0039】画像入力部 21 は、図 1 において説明した画像入力部 11 と同じものとして説明を省略する。

【0040】2 値化処理部 23 は、画像入力部 23 から取り込まれた文書画像を、公知である 2 値化処理により白と黒の 2 値の画像データに変換し、前処理部 25 に出力する。

【0041】前処理部 25 は、2 値化処理部 23 から出力された 2 値画像について、例えば文献「信学技報、P R U 92-32、1992」に記載されている傾き検出・補正処理により、傾きのない 2 値画像に変換する。さらに、前処理部 25 は、傾きが補正された 2 値画像に対して、公知である黒連結成分抽出処理により、連結する黒画素のまとまりを囲む外接矩形枠を生成し、その大きさ（縦幅と横幅の長さ）や位置座標値を抽出、管理し、その結果を特徴抽出部 27 に出力する。

【0042】なお、入力文書の位置座標は左上端を原点とし、x 座標値は右方向に次第に大きくなり、y 座標値は下方向に次第に大きくなるように定義されているものとする。本実施例で触れる文書画像は全てこの座標系で定義されている。この内、縦幅と横幅の長さが、それぞれ動的に検出されたしきい値 th_1 、 th_2 よりも小さく、かつ最近傍の他の黒連結成分矩形から距離 th_3 以上離れているものを、ノイズあるいは網点であるとして除去するようにしてもよい。

【0043】以後、画像入力部 21 から入力された文書画像に対して、2 値化処理部 23 における 2 値化処理、前処理部 25 における前処理を施して得た画像を入力文書画像または単に入力画像と呼ぶ。入力画像は、特徴抽出部 27 に送られる特徴抽出部 27 は、前処理部 25 から得られた入力画像から幾何学的な図形特徴を抽出するものであり、図 4 に示すように、線分抽出部 27a、文字候補矩形抽出部 27b、及び罫線特徴抽出部 27c によって構成されている。特徴抽出部 27 は、以下に示す

手順で幾何学的な図形特徴を抽出する。

【0044】まず、線分抽出部27aは、例えば、以下に述べる手順に基づいて入力画像から線分を抽出する。このとき線分抽出処理は、垂直方向と水平方向の2つの方向に限定して実施されるようにしてもよい。具体的な例として、垂直方向の線分を抽出する場合について説明する。なお、水平方向の線分についても同様の手順により抽出することができる。

【0045】Step1: 文書画像を垂直方向に順次走査し、各走査線において、予め定めた例えば3ドット以上の長さを持つ連続する黒画素の連なり集合($BL = \{b_{li} \mid i=1, 2, \dots, n\}$)を抽出する。

【0046】Step2: BLの要素のうち、水平方向に隣接しているものを統合してまとめることにより、さらに、その集合($BLG = \{b_{lgj} \mid j=1, 2, \dots, M\}$)を抽出する。

【0047】Step3: BLGの各要素に含まれるBLにおいて、垂直方向に最も長い黒画素の連なり b_{lmax} を抽出し、その長さ b_{lmax} の α (例えば $\alpha=0.3$)倍未満の長さを持つ b_{li} を削除する。

【0048】Step4: 残った b_{li} を内接する矩形を抽出し、その集合($RL = \{r_{lk} \mid k=1, 2, \dots, P\}$)を抽出する(このとき、 r_{lk} とそれを構成する b_{li} とを相互に関連づける)。 r_{lk} の左上端および右下端のy座標値を検出し、それぞれ r_{lky1} 、 r_{lky2} とする。

【0049】Step5: RLの各要素を水平方向(左から右への方向)に順次走査し、各走査線において最初に出現する b_{li} のx座標(x_{is})と最後に出現する b_{li} のx座標(x_{ie})を保持する。

【0050】Step6: 各 r_{lk} の各水平走査線において得られた x_{is} の集まりと、 x_{ie} の集まりの平均値 r_{lkx1} 、 r_{lkx2} をそれぞれ計算する。

【0051】Step7: 各 r_{lk} の左上端および右下端の座標をそれぞれ(r_{lkx1} 、 r_{lky1})、(r_{lkx2} 、 r_{lky2})に設定し、それを線分として抽出する。

【0052】線分抽出部27aにより、入力画像から抽出された線分の例を図5に示している。図5に示すように、この段階では文字の部分において、文字を構成する短い線分を検出している。これらを取り除くように以下の処理を行なってもよい。すなわち、線分抽出部27aによる線分抽出処理に前後して、あるいは同時に黒連結成分群に対して、文字候補矩形抽出部27bは、以下に述べる手順により、文字とみなすことのできる黒連結成分(以後、文字候補矩形と呼ぶ)を選出する。

【0053】Step1: 各黒連結成分を内接する矩形を抽出し、その縦幅 ch と横幅 cw を求める。

【0054】Step2: ch と cw の各値に対して出現頻度を求め、最頻値を抽出する。最頻値を示す ch を

入力文書中の文字の平均的な文字高さ(CH)、最頻値を示す cw を入力文書中の文字の平均的な文字幅(CW)とみなす。

【0055】Step3: $(CW - th) \leq cw \leq (CW + th)$ かつ $(CH - th) \leq ch \leq (CH + th)$ を満たす黒連結成分を文字候補矩形として抽出する。

【0056】文字候補矩形抽出部27bにより抽出された文字候補矩形の抽出結果例を図6(文字部を外枠矩形で囲んだ例)に示している。

【0057】なお、文字候補矩形抽出処理を線分抽出処理より先に行ない、得られた文字候補矩形以外を入力画像上の部分に対して垂直方向と水平方向の両方向に、以下に述べるフィルタリング処理を適用して、罫線がかすれていたり、とぎれているような場合でも線分抽出処理結果を安定させるようにしてもよい。

【0058】この場合、特徴抽出部27は、図7に示すように構成される。すなわち、文字候補矩形抽出部27bの処理結果がフィルタリング処理部27dに出力されフィルタリング処理が施される。そして、フィルタリング処理の後に、線分抽出部27aによる線分抽出処理実行される。

【0059】フィルタリング処理部27dによるフィルタリング処理は、例えば、2値画像において水平(垂直)方向の走査線上のある決まった長さ以内で連続する白画素を全て黒画素に置き換えるという処理で実現される。これにより、かすれやとぎれのある罫線部分が補正されるので、線分抽出処理結果が安定される。

【0060】罫線特徴抽出部27cは、線分抽出部27aによって抽出された線分群において、図8に示すような文字候補矩形に交差あるいは包含されている線分を除去し、残った線分を罫線特徴として抽出する。

【0061】この時点では、各罫線特徴は、入力画像における当該画像を囲む外接矩形の左上端と右下端の位置座標、外接矩形の縦幅及び横幅などの種々の情報で表現されるようにしてもよい。また、罫線特徴に「対応付いた罫線特徴の識別番号」を用意しておいて、後段のモデル照合部35で実施される照合処理において、対応付いた相手の識別番号を格納するようにしてもよい。

【0062】得られた罫線特徴の集合は、さらに水平罫線と垂直罫線の2種類に分類され、水平罫線特徴の集合とその要素数、垂直罫線特徴の集合とその要素数が抽出される。

【0063】入力画像に対する罫線特徴抽出結果の例を図9に示す。ここで、入力画像から抽出された水平罫線の集合(水平罫線特徴集合)と垂直罫線の集合(垂直罫線特徴集合)の各要素に、

$IHL = (ihl1, ihl2, ihl3, ihl4, ihl5, ihl6, ihl7),$

$IVL = (ivl1, ivl2, ivl3, ivl4, ivl5, ivl6, ivl7, ivl8, ivl9)$

となるように識別番号を付与する。

【0064】また、水平罫線特徴集合の要素数を $i h\text{-}num$ 、垂直罫線特徴集合の要素数を $i v\text{-}num$ とし、水平罫線特徴集合と垂直罫線特徴集合をまとめて入力罫線特徴集合と呼ぶことにする。入力罫線特徴集合は、後段の特徴構造化部 29 に送られる。以後の説明において上記記号を用いることにする。

【0065】特徴構造化部 29 は、特徴抽出部 27 から得られた入力罫線特徴集合から以下に述べる手順で罫線に関する構造化特徴を抽出するもので、図 10 に示すように、罫線グループ化処理部 29 a、表特徴抽出部 29 b、罫線接合部検出部 29 c、及び特徴間関係記述部 29 d によって構成されている。

【0066】まず、罫線グループ化処理部 29 a は、交差・接続する罫線特徴を一まとめにグループ化する。罫線グループ化処理アルゴリズムは例えば以下になる。このアルゴリズムは、交差・接続する罫線には同じラベルを与えることにより罫線をグループ化するものである。

【0067】Step 1: ラベル番号を初期化する。

【0068】Step 2: 1 本の垂直罫線を選択する。

【0069】Step 3: ステップ 2 で選択された当該垂直罫線に直交する全ての水平罫線を抽出する。

【0070】Step 4: 当該垂直罫線か当該水平罫線のいずれかに既にラベルが付与されている場合には、その中で最小値のラベルを当該罫線の全てに付与する。

【0071】Step 5: 当該垂直罫線と当該水平罫線のいずれにもラベルが付与されていない場合には、それら全てに新しいラベル番号を付与し、ラベル番号を更新する。

【0072】Step 6: Step 2 から Step 5 を全ての垂直罫線に適用する。(Step 2 から Step 6 までを手続き A とする)

Step 7: 1 本の水平罫線を選択する。

【0073】Step 8: 当該水平罫線に直交する全ての垂直罫線を抽出する。

【0074】Step 9: 当該水平罫線か当該垂直罫線のいずれかに既にラベルが付与されている場合にはその

表特徴: $i t_1$

左上端の位置座標

右下端の位置座標

表の縦幅

表の横幅

重心の位置座標

水平罫線数

当該表に含まれる水平罫線の集合

垂直罫線数

当該表に含まれる垂直罫線の集合

当該表に含まれる接合部数

他の表に内接されているか否かの情報

中で最小値のラベルを当該罫線の全てに付与する。

【0075】Step 10: 当該垂直罫線と当該水平罫線のいずれにもラベルが付与されていない場合には、それら全てに新しいラベル番号を付与し、ラベル番号を更新する。

【0076】Step 11: Step 7 から Step 10 を全ての垂直罫線に対して適用する (Step 7 から Step 11 までを手続き B とする)。

【0077】Step 12: 手続き A と手続き B をラベル番号の更新がなくなるまで繰り返す。

【0078】Step 13: 同じラベルが付与されている罫線特徴をグループ化する。

【0079】例えば、図 9 に示す入力罫線特徴集合に対して、前述のようなグループ化処理を適用すると以下に示す 2 つのグループが得られる。

【0080】Group 1: ($i h 1_1, i h 1_2, i h 1_3, i h 1_4, i h 1_5, i v 1_1, i v 1_2, i v 1_3, i v 1_4, i v 1_5, i v 1_6$),

Group 2: ($i h 1_6, i h 1_7, i v 1_7, i v 1_8, i v 1_9$)

表特徴抽出部 29 b は、罫線グループ化処理部 29 a によって得られたグループごとに、グループに含まれる罫線を内接する矩形 (以後、表枠と呼ぶ) を抽出する。さらに、表特徴抽出部 29 b は、表特徴として、例えば左上端の位置座標、右下端の位置座標、表の縦幅、表の横幅、重心の位置座標、水平罫線数、当該表に含まれる水平罫線の集合、垂直罫線数、当該表に含まれる垂直罫線の集合、当該表に含まれる接合部数、他の表に内接されているか否かの情報などを抽出する。この時、水平 (垂直) 罫線特徴は、その左上端の $y (x)$ 座標値の昇順にソートされていてもよい。

【0081】ここで接合部数については、後段の罫線接合部検出部 29 c で抽出され、表特徴として格納される。例えば、Group 1 からは、図 11 に示すように以下の表特徴 $i t_1$ が得られる。また、Group 2 から得られる表特徴を $i t_2$ とする。

【0082】

($i t_1 x_1, i t_1 y_1$)

($i t_1 x_2, i t_1 y_2$)

$i t_1 height$

$i t_1 width$

($i t_1 c x, i t_1 c y$)

$i h 1_1 num$

$I H L_1$

$i v 1_1 num$

$I V L_1$

$i j_1 junc\text{-}num$

$nest flag$

$it1\ height = it1\ y2 - it1\ y1 + 1,$
 $it1\ width = it1\ x2 - it1\ x1 + 1,$
 $it1\ cx = (it1\ x1 + it1\ x2) / 2,$
 $it1\ cy = (it1\ y1 + it1\ y2) / 2,$
 $ih1\ num = 5,$
 $IHL1 = (ih11, ih12, ih13, ih14, ih15),$
 $iv1\ num = 6,$
 $IVL1 = (iv11, iv12, iv13, iv14, iv15, iv16),$
 $nest\ flag = 0$ (すなわち他の表に含まれていない)
 1枚の入力帳票に複数の表が含まれていることを考慮して、さらにページ特徴を抽出、管理する。ページ特徴は、例えば、表数、表特徴の集合、水平罫線数、垂直罫線数、及び接合部数によって定義される。
 【0083】例えば、図9の入力罫線集合からは、ページ特徴：IPとして、
 表数=2、
 表特徴の集合IT=(it1, it2)、
 水平罫線数ih1-num=7、垂直罫線数iv1-num=9、接合部数ijunc-num=28、が得られる。
 【0084】次に、罫線接合部検出部29cは、各表において、そこに含まれる水平罫線と垂直罫線の交差部分・接続部分(以後、接合部と呼ぶ)を求め、さらに各表特徴で接合部の個数を管理する。罫線接合部検出部29cの動作手順は例えば次のようになる。
 【0085】Step1: 一本の水平罫線を選択する。
 【0086】Step2: 着目した水平罫線に交差・接続する全ての垂直罫線を抽出する。
 【0087】Step3: 当該水平罫線と当該垂直罫線の交点を求め、水平罫線ごとにその座標値を管理する。各水平罫線では接合部特徴はx座標値の昇順にソートしておく。
 【0088】Step4: Step1からStep3までを全ての水平罫線に対して実施する。
 【0089】Step5: 一本の垂直罫線を選択する。
 【0090】Step6: 着目した垂直罫線に交差・接続する全ての水平罫線を抽出する。
 【0091】Step7: 当該垂直罫線と当該水平罫線の交点を求め、垂直罫線ごとにその座標値を管理する。各垂直罫線では接合部特徴はy座標値の昇順にソートしておく。
 【0092】Step8: Step5からStep7までを全ての垂直罫線に対して実施する。
 【0093】例えば、表特徴it1では、水平罫線集合IHL1と、垂直罫線集合IVL2から、図12に示す接合部が得られ、その接合部数「22」を表特徴に加える(ijunc-num=22とする)。
 【0094】特徴間関係記述部29dは、以上の処理結果より得られた情報を、例えば図13のように関係づけ

て管理する。この結果、ページ特徴から表特徴、罫線特徴、接合特徴と階層的に関連づけられて管理され、特徴に関する情報を効率的に検索できるようになる。

【0095】以後、これらの特徴を総称して構造化特徴と呼ぶ。また、入力文書から抽出された構造化特徴を入力構造化特徴、モデル文書から抽出された構造化特徴をモデル構造化特徴と呼ぶ。

【0096】一方、処理対象文書に対する処理とは別に、モデル登録部31は、オペレータによって提示された処理対象文書の種類ごとの一例であるサンプル文書をもとに、オペレータとの間の対話的入力作業によりモデルの登録を行なう。ここでは、モデル登録作業時のモデル登録部31の動作の一例について説明する。

【0097】まず、画像入力部21を介して登録対象のサンプル文書を入力し、2値化処理部23、前処理部25を経て文書画像(以後、モデル画像と呼ぶ)を取得する。次いで、特徴抽出部27においてモデル画像から罫線特徴を抽出する。

【0098】モデル登録部31は、特徴抽出部27における抽出結果に対して特徴抽出処理の誤りを推定し、それを修正する旨のメッセージを、表示部31aのディスプレイの画面上に表示して、修正作業をオペレータに指示する。

【0099】特徴抽出処理の誤りの推定方式は、例えば、端点が接合部となっていない罫線を見つけて、その罫線が正しく抽出されていないとみなすことにより実現できる。モデル登録部31は、表示部31aに表示された指示に従いオペレータによって画面上で修正された罫線特徴を入力部31bを介して入力する。

【0100】修正作業が完了すると、修正された罫線特徴は特徴構造化部29に送られ、構造化特徴が抽出される。この構造化特徴は、未知入力文書进行处理する際に、フォーマット種別同定部33、モデル照合部35、未対応・矛盾対応発見修正部39、照合結果判定部37などで用いられる。

【0101】抽出・修正された罫線特徴は、画面上にモデル画像と重ねて表示される。オペレータは、書式構造化情報登録部31が画面上に出力するメッセージに従いながら、モデル文書に対応付くべき入力文書画像进行处理するために必要な知識を構築していく。

【0102】この結果、モデルとして構造化特徴の情報、文字認識対象領域の指定に関する情報、文字認識対象領域の文字列方向の情報、文字読み取り処理時に拘束条件として働く文字種情報、筆記形態情報、文字認識後処理のための項目属性情報、文字認識結果の対応関係などの情報が知識ベース31cに格納される。

【0103】このうち構造化特徴は、正立した画像に対して作られている。そこで、モデル登録部31は、知識ベース31cへの登録時に、左に90度回転したものと右に90度回転したものと180度回転したものをそれ

ぞれ発生させ、さらに何度回転されているかを示す情報を付加するようにしてもよい。

【0104】これら4方向の構造化特徴と入力構造化特徴は、後段のモデル照合部35において、それぞれの方向の特徴と入力画像とを照合させることにより、入力文書の文書方向が未知であっても、モデルマッチングで最良マッチングした構造化特徴に付与されている回転角度を知ることができる。従って、その角度を0度にする方向に入力画像を回転させれば、必ず正立した入力画像を得ることができる。

【0105】モデル登録部31は、以上の処理を全てのモデル文書に対して行なう。ここでは、一例として図14と図15に示す2種類の構造化特徴がモデルとして登録されたものとする（以後、図14をモデル1、図15をモデル2と呼ぶ）。

【0106】フォーマット種別同定部33は、モデル登録部31によって予め登録されているモデルから、入力文書に対応するもの（あるいは対応する可能性のあるもの）を選出する。

【0107】ここでは、入力文書の構造化特徴と登録されている全てのモデルの構造化特徴との間で類似度計算を行ない、最も類似度の高いモデルあるいは、ある一定の値以上の類似度を有するモデルの構造化特徴を後段のモデル照合部35に送り込む。

【0108】類似度計算に用いる特徴としては、例えば文書中に含まれる、特徴構造化29によって抽出された接合部数を用いる。接合部数という特徴は、未知入力文書が寸法の拡大・縮小、さらには各表が独立してスケール変換されている場合にも、影響を受けない特徴であるので書式構造種別同定のための類似度計算に適している。

【0109】この他に、文書中に含まれる水平罫線特徴数や垂直罫線特徴数、表特徴数なども類似度計算のための特徴として有効である。ここでモデル文書は適宜に順序づけがなされているとする。

【0110】また、類似度は以下のようにして求まる。ここで、モデルの総数： $model_num$ 、入力文書中の接合部数： i_j_n 、 k 番目のモデル文書中の接合部数： $m_j_n_k$ （ただし $1 \leq k \leq model_num$ ）、入力文書と k 番目のモデル文書との類似度： $s_i_m_k$ （ただし $-100 \leq s_i_m_k \leq 100$ ）とすると、類似度は例えば、
 $i_j_n > 0$ または $m_j_n_k > 0$ のとき

$$s_i_m_k = 100 - 200 \times |i_j_n - m_j_n_k| / (i_j_n + m_j_n_k)$$

 $i_j_n = 0$ かつ $m_j_n_k = 0$ のとき
 $s_i_m_k = -100$
 により求まる。

【0111】得られた類似度のうち類似度が最大であるもの、あるいは予め設定されたしきい値 $th5$ 以上のものを候補のモデルとして入力文書の構造化特徴とともにモ

デル照合部35に送る。このとき、全てのモデルにおいて類似度がしきい値 $th5$ 未満である場合、以後の処理を中断して、入力文書を棄却してもよいし、また、全てのモデルの構造化特徴をモデル照合部に送り込むようにしてもよい。

【0112】例えば、登録されているモデル文書が図14と図15の2種類である場合、入力文書とそれら2つのモデル文書との類似度計算は以下ようになる。ここで入力文書の接合部数： $i_j_n = 28$ 、モデル1の接合部数： $m_j_n_1 = 31$ 、モデル2の接合部数： $m_j_n_2 = 26$ とし、入力文書とモデル1との類似度を $s_i_m_1$ 、入力文書とモデル2との類似度を $s_i_m_2$ 、しきい値： $th5 = 80$ とする。

【0113】

$$s_i_m_1 = 100 - 200 \times |28 - 31| / (28 + 31) = 89.83$$

$$s_i_m_2 = 100 - 200 \times |28 - 26| / (28 + 26) = 92.59$$

以上の計算結果から $s_i_m_1$ と $s_i_m_2$ が共にしきい値 $th5$ を越えているので、双方とも入力文書に対応付く可能性のあるモデルとしてそれらの構造化特徴を後段のモデル照合部35に送る。

【0114】モデル照合部35は、入力文書から抽出された構造化特徴（以後、入力構造化特徴と呼ぶ）と、書式構造種別同定部で選出された一つあるいは複数のモデル文書の構造化特徴（以後、モデル構造化特徴と呼ぶ）を受け取り、それらの間で以下に述べる照合処理を行ない、照合度を計算する。

【0115】照合処理は、当該入力構造化特徴と当該モデル構造化特徴との間の対応付け処理のことであり、照合度はその対応付きの程度を表す尺度である。ここでは、最も高い照合度を示すモデル文書が入力文書に対応するものであるとし、入力構造化特徴とそのモデル構造化特徴との間の対応関係に関する情報を、後段の照合結果判定部に送り込む。最大照合度を示すモデル文書と入力文書との間の対応関係が獲得されているか否かの最終的な判断は照合結果判定部で行なわれる。

【0116】特徴構造化部29で得られた構造化特徴は、図13に示すように、ページ特徴から表特徴へ、さらに表特徴から罫線特徴へと全体一部分という階層的な関係が抽出されて構造化されている。この特質を用いる場合、照合処理は階層的に実施される。これに対応して、モデル照合部35は、さらに図16に示すように、選択部35a、表照合部35b、罫線照合部35c、照合度計算部35d、及び照合結果出力部35eによって構成される。

【0117】まず、選択部35aは、複数のモデルの構造化特徴から一つのモデルの構造化特徴を選択し、入力文書の構造化特徴とともに表照合部35bに送る。

【0118】次いで、表照合部35bは、入力文書の表

特徴の集合（以後、入力表特徴集合と呼ぶ）とモデル文書の表特徴の集合（以後、モデル表特徴集合と呼ぶ）との間で対応付けを行なう。

【0119】表間対応がとれた場合には、さらに罫線照合部35cは、入力文書の当該表の内部に含まれる罫線特徴（以後、入力罫線特徴と呼ぶ）とモデル文書の当該表の内部に含まれる罫線特徴（以後、モデル罫線特徴と呼ぶ）との間で対応付けを行なう。

【0120】罫線間の対応がとれた場合には、照合度計算部35dは、（入力文書と当該モデル文書との間の）照合度を計算する。ここで、表間対応と罫線間対応のどちらも得られなかった場合には、照合度に最低値（-100）を代入する。

【0121】照合結果出力部35eは、それまでに計算されている最も高い照合度を示すモデル文書と入力文書の組を選び、その構造化特徴間の対応関係に関する情報を後段の照合結果判定部37に出力する。

【0122】以下では、フォーマット種別同定部33から図14と図15の2種類のモデルと入力文書の構造化特徴がそれぞれ送られてきた場合のモデル照合部35の処理動作例について説明する。

【0123】2種類のモデルのうち、選択部35aによってモデル1が選ばれ、入力文書とともにそれぞれの構造化特徴が表照合部35bに送り込まれたとする。ここで、入力文書の表特徴数を $i\text{-num}=2$ とし、表特徴集合を IT とすると、 $IT=(it_1, it_2)$ （図9参照）モデル1の表特徴数を $mt\text{-num}=2$ とし、表特徴集合を $MT1$ とすると、 $MT1=(mt_{11}, mt_{12})$ （図14参照）とする。

【0124】表照合部35bは、さらに図17に示すように、対応可能ペア検出部35b-1、異種対応可能ペア間両立関係判定部35b-2、及び最良マッチング抽出部35b-3によって構成されており、次のように動作する。

【0125】まず、対応可能ペア検出部35b-1は、選択部35aで選択されたモデルの表特徴集合の各要素に対して、それと対応付く可能性のある入力の特徴をすべて検出し、対応可能ペアとして管理する。すなわち、

Step1: 任意の MT から任意の一つの表特徴 mt_k （着目モデルの表特徴集合のうち k 番目の表特徴）を選ぶ。

【0126】Step2: 表特徴 mt_k の接合部数と IT の各表特徴の接合部数との間で類似度を計算をする。

【0127】Step3: 例えば mt_k と it_j （入力表特徴集合の j 番目の表特徴）の類似度 sim_{kj} が、あらかじめ設定したしきい値 $th6$ 以上である場合には、対応可能であるとして、その組 (mt_k, it_j) を対応可能ペアとして保持する。

【0128】Step4: Step3を表特徴集合 IT

のすべての要素に対して適用する。

【0129】Step5: Step1~Step4までを着目 MT のすべての要素に対して適用する。

【0130】ここで、 it_j の接合部数を ijn_j 、 mt_k の接合部数を mjn_k とすると、類似度は、例えば次の式で求まるとする。

【0131】 $ijn_j > 0$ または $mjn_k > 0$ のとき、 $sim_{kj} = 100 - 200 \times |ijn_j - mjn_k| / (ijn_j + mjn_k)$

$ijn_j = 0$ かつ $mjn_k = 0$ のとき、

$sim_{kj} = -100$

上記動作を IT と $MT1$ を例として具体的に説明すると以下ようになる。 $th6 = 75$ 、 it_1 の接合部数=22、 it_2 の接合部数=6、 mt_{11} の接合部数=25、 mt_{12} の接合部数=6とすると式から、

mt_{11} と it_1 の間の類似度： $sim_{111} = 86.67$ となり、 $sim_{111} > th6$ であるから、対応可能であるとして、その組 (mt_{11}, it_1) を対応可能ペアとして保持する。

【0132】 mt_{11} と it_2 の間の類似度： $sim_{112} = -22.58$ となり、 $sim_{112} < th6$ であるから、対応不可能であるとする。

【0133】 mt_{12} と it_1 の間の類似度： $sim_{121} = -14.28$ となり、 $sim_{121} < th6$ であるから、対応可能であるとする。

【0134】 mt_{12} と it_2 の間の類似度： $sim_{122} = 100$ となり、 $sim_{122} > th6$ であるから、その組 (mt_{12}, it_2) を対応可能ペアとして保持する。

【0135】以上の結果、対応可能ペアとして、 IT と $MT1$ の間では、 $p1 = (mt_{11}, it_1)$ と $p2 = (mt_{12}, it_2)$ が検出された。

【0136】次に、異種対応可能ペア間両立関係判定部35b-2は、当該モデル文書と入力文書の間で、対応可能ペア検出部35b-1で検出された2つの異なる対応可能ペアが両立するものであるか否かを判定する。

【0137】ここでいう「両立する」ということは、2つの対応可能ペアが同時に存在することに矛盾が無いことを意味する。ここでの判定条件としては以下のものが上げられる。判定対象となる対応可能ペアをそれぞれ (mt_k, it_j) 、 (mt'_k, it'_j) とする。

ここで、 mt_k の左上端の位置座標を $(mt_k x1, mt_k y1)$ 、右下端の位置座標を $(mt_k x2, mt_k y2)$ 、 mt'_k の左上端の位置座標を $(mt'_k x1, mt'_k y1)$ 、右下端の位置座標を $(mt'_k x2, mt'_k y2)$ 、 it_j の左上端の位置座標を $(it_j x1, it_j y1)$ 、右下端の位置座標を $(it_j x2, it_j y2)$ 、 it'_j の左上端の位置座標を $(it'_j x1, it'_j y1)$ 、右下端の位置座標を $(it'_j x2, it'_j y2)$ とする。

【0138】また、文書画像の座標系 (x, y) ： $0 <$

$x \leq \text{WIDTH}$, $0 < y \leq \text{HEIGHT}$ において、図18に示すような座標系を定義する。すなわち、図18(a)中に示す矩形領域T: (左上端の位置座標を(t_{x1} , t_{y1})、右下端の位置座標を(t_{x2} , t_{y2})とする)に対して、図18(b)に示す $0 < x < t_{x1}$ かつ $0 < y \leq \text{HEIGHT}$ を満たす領域を領域1、図18(c)に示す $t_{x2} < x \leq \text{WIDTH}$ かつ $0 < y \leq \text{HEIGHT}$ を満たす領域を領域2、図18(d)に示す $0 < x \leq \text{WIDTH}$ かつ $0 < y < t_{y1}$ を満たす領域を領域3、図18(e)に示す $0 < x \leq \text{WIDTH}$ かつ $t_{y2} < y \leq \text{HEIGHT}$ を満たす領域を領域4と定義する。

【0139】また、判定条件として、以下の条件1から条件3までのすべての条件を満たさない場合のみ、2つの対応可能ペアが両立可能であると見なす。

【0140】条件1: $mt_k = mt'_k$ 、

条件2: $it_j = it'_j$ 、

条件3: 配置関係に逆転があること。条件3は、以下の条件3-1~3-4の条件を満たす場合である。すなわち、

条件3-1: it_j (it'_j) に対して、 it'_j (it_j) が領域1にあり、 mt_k (mt'_k) に対して、 mt'_k (mt_k) が領域2にある。

【0141】条件3-2: it_j (it'_j) に対して、 it'_j (it_j) が領域2にあり、 mt_k (mt'_k) に対して、 mt'_k (mt_k) が領域1にある。

【0142】条件3-3: it_j (it'_j) に対して、 it'_j (it_j) が領域3にあり、 mt_k (mt'_k) に対して、 mt'_k (mt_k) が領域4にある。

【0143】条件3-4: it_j (it'_j) に対して、 it'_j (it_j) が領域4にあり、 mt_k (mt'_k) に対して、 mt'_k (mt_k) が領域3にある。

【0144】異種対応可能ペア間両立関係判定部35b-2の動作を、前述した対応可能ペア $p1$ と $p2$ を用いて具体的に説明する。 IT と $MT1$ の間における2つの異なる対応可能ペアの組として($p1$, $p2$)がある。このペアは条件1から条件3までのすべてを満たさないで両立可能であると判断される。異種対応可能ペア間両立関係判定部35b-2は、両立可能と判断された対応可能ペアを、最良マッチング抽出部35b-3に出力する。

【0145】最良マッチング抽出部35b-3は、異種対応可能ペア間両立関係判定部35b-2において両立すると判定された対応可能ペアのうち、すべてが互いに両立可能な対応可能ペアの最大の集合を求めることにより、入力表特徴集合とモデル表特徴集合間の最良マッチングを抽出する。

【0146】最良マッチング抽出部35b-3は、さら

に図19に示すように、連合グラフ作成部35b-3aと最大クリーク抽出部35b-3bにより構成されている。連合グラフ作成部35b-3aは、異種対応可能ペア間両立関係判定部35b-2の判定結果を受け取り、その情報をもとに連合グラフを作成する。この連合グラフの節点は対応可能ペアを示し、節点間を結ぶ弧は2つの対応可能ペアが両立可能であることを示すものであり、異種対応可能ペア間両立関係判定部35b-2の判定結果を表す補助的なデータ構造である。

【0147】具体的には、対応可能ペア $p1$ と $p2$ が連合グラフ作成部35b-3aに送られ、この結果、図20に示す連合グラフが得られる。上述した「すべてが互いに両立可能な対応可能ペアの最大の集合」はこの連合グラフにおいては全体的に連結した(完全に互いに両立可能な)最大の節点集合であると捉えることができる。それはクリークであり、本実施例ではより大きなクリークはより良いマッチングを表している。

【0148】最大クリーク検出部35b-3bは、連合グラフ作成部35b-3aによって作成された連合グラフから、節点数が最大である完全連結部分グラフ(最大クリーク)を抽出する。なお、最大クリーク検出部35b-3bの動作は、例えば文献「信学論(D), J68-D, 3, pp221-228, 1985」に記載されている方式を適用することができる。

【0149】最大クリーク検出部35b-3bによって抽出されたクリーク、すなわち、「すべてが互いに両立可能な対応可能ペアの最大の集合」を構成する要素の対応可能ペアは、入力文書の表特徴集合 IT と当該モデル文書の表特徴集合 $MT1$ との間に対応の取れた表の組を表す。具体的には、最大クリーク: $mclicue = p1, p2$ が得られ、 mt_{11} と it_1 、 mt_{12} と it_2 の対応がそれぞれ得られたことになる。この対応結果は、後段の罫線照合部35cに送られる。

【0150】このとき、入力文書とモデル文書の表特徴集合の各表では、その内部に含まれる罫線特徴の座標値は、表の左上端の座標値で正規化されている。また、入力文書の各表の罫線特徴の座標値は、対応付くモデル文書の表に重ね合わせられるようにスケール変換されている。

【0151】すなわち、入力表の縦幅が $it\text{-height}$ 、横幅が $it\text{-width}$ 、対応付いたモデル表の縦幅が $mt\text{-height}$ 、横幅が $mt\text{-width}$ であるとき、当該入力表の正規化された罫線特徴の座標値は、さらに、 x 座標値では $(mt\text{-width} / it\text{-width})$ 倍、 y 座標値では $(mt\text{-height} / it\text{-height})$ 倍されることによりスケール変換される。この結果、後段の罫線照合部35cでは同じスケールかつ同じ座標系のもとで入力表とモデル表に含まれる罫線集合間の対応付け処理が可能になる。

【0152】前述した対応可能ペア検出部35b-1の動作を示すStep3において、 $simkj < th6$ である

場合には、 mt_k と it_j の構造が異なっている他に、例えば図 21 (a) に示すように、入力文書の印刷品質が悪いために表が分離している場合 (図中 (a-1)) や、隣接する表の間隔が狭いために画像入力時に接触してしまったりする場合 (図中 (b-1)) も考えられる。

【0153】このような問題点に対応するために対応可能ペア検出部 35b-1 は、さらに以下の処理を実施するようにしてもよい。

【0154】Step 3-1: $m_{jnk} > i_{jn}$ ($m_{jnk} < i_{jn}$) である場合には、IT (MT1) の中から以下の条件を満たす it_j (mt_k) に隣接する表特徴 it_l (mt_l) を 1 つ検出し、それらを統合して仮想的な表特徴 it'_j を新たに生成し、その際に生じる接合部数 $i_{jn'}$ ($m_{jn'}$) と m_{jnk} (i_{jn}) との類似度 $siml'_{kj}$ を計算する。

【0155】類似度: $siml'_{kj}$ がしきい値 $th6$ 以上である場合には、対応可能であるとして、(mt_k, it'_j) (mt'_k, it_j) を対応可能ペアとして保持する。

【0156】条件: it_j (mt_k) と it_l (mt_l) の間に他の表が存在しないこと。

【0157】Step 3-2: Step 3-1 で求めた $siml'_{kj}$ がしきい値 $th6$ に満たない場合、類似度: $siml'_{kj}$ がしきい値 $th6$ 以上となるまで、あるいは条件を満たす統合すべき表が見つからなくなるまで Step 3-1 を繰り返す。

【0158】この場合、表照合部 35b の最良マッチング部 35b-3 では、最大クリークを構成する「対応付き」のすべての組み合わせを出力するようにして、その後次に次の条件を適用することにより候補を絞り込む。それでもなお複数の組み合わせが生じている場合には、それらすべてを後段の罫線照合部 35c に送り込み、その結果に基づいて (それらの中で照合度が最も高くなる組み合わせを選ぶことにより) 最終的に表間対応関係を一意に決めるようにしてもよい。

【0159】条件: モデル表特徴集合のすべての要素が入力表特徴集合のいずれかの要素に対応していること。

【0160】次に、罫線照合部 35c では、表照合部 35b で対応付いた入力表とモデル表にそれぞれ含まれる罫線集合間で対応付け処理を行なう。このとき水平罫線集合間の対応付けと垂直罫線集合間の対応付けを独立に行なうようにしてもよい。そして両者の整合を、各々の対応付け処理が済んだあとで得るようにしてもよい。この場合の罫線照合部 35c は、さらに図 22 に示すように構成される。すなわち、表対応選択部 35c-1、垂直罫線照合部 35c-2、水平罫線照合部 35c-3、及び方向間整合獲得部 35c-4 によって構成されている。

【0161】まず、表対応選択部 35c-1 は、表照合

部 35b で抽出された表間対応付きの中から任意の対応を選択する。垂直罫線照合部 35c-2 は、表対応選択部 35c-1 によって選択された入力表とモデル表の垂直罫線集合間で対応付けを行なう。この対応付けに成功すれば、さらに水平罫線照合部 35c-3 は、水平罫線集合間で対応付けを行なう。方向間整合獲得部 35c-4 は、垂直罫線照合部 35c-2 による対応付けと水平罫線照合部 35c-3 による対応付けとの間の整合を獲得する。

【0162】なお、垂直罫線照合部 35c-2 と水平罫線照合部 35c-3 は、さらに図 23 に示すように構成されている。これらの処理動作は、罫線方向の違いを考慮する以外は、基本的に同じである。

【0163】以下に、具体的な説明として、垂直罫線照合部 35c-2 における、 mt_{11} の垂直罫線集合 $M1VL1$ と it_{11} の垂直罫線集合 $IVL1$ の間の対応付け処理の動作について説明する。ただし、図 14 より、 $M1VL1 = (m1v11, m1v12, m1v13, m1v14, m1v15, m1v16, m1v17, m1v18)$ とする。以後、垂直罫線を単に罫線と略して説明する。

【0164】図 23 中に示す対応可能罫線特徴ペア検出部 35c-2a は、モデルの罫線集合の各要素に対して、要素と対応付く可能性のある入力罫線特徴をすべて検出し、対応可能な罫線特徴のペアとして管理する。すなわち、以下に説明する手順を実行する。

【0165】Step 1: 任意の $M1VL1$ から任意の一つの罫線特徴 $m1v1k$ (k 番目の罫線特徴を意味する) を選ぶ。

【0166】Step 2: $m1v1k$ に対応付く入力罫線特徴を検出するための探索範囲 ($areamv1$) を設定する。

【0167】ここで、 $m1v1k$ の罫線特徴の左上端と右下端の位置座標を ($m1x1, m1y1$)、($m1x2, m1y2$) とすると、探索範囲は例えば ($m1x1 - th9, m1y1$) と ($m1x2 + th9, m1y2$) の座標値で構成される矩形の内部としてもよい。ここで、しきい値 $th9$ が予め与えられているものとする。

【0168】Step 3: Step 2 で求められた探索範囲に内包・交差する入力罫線特徴を抽出する。

【0169】この処理は例えば以下のようにして行なわれる。すなわち、探索対象となる入力罫線の左上端と右下端の位置座標値を ($ix1, iy1$)、($ix2, iy2$) とした場合、以下の条件を満たす入力罫線を抽出する。

【0170】条件: $\min(m1x2 + th9, ix2) - \max(m1x1 - th9, ix1) + 1 > 0$ かつ $\min(m1y2, iy2) - \max(m1y1, iy1) + 1 > 0$ である。

【0171】Step 4: 抽出された入力罫線特徴の一

つを選ぶ(例えば、 $ivlj$ (入力罫線特徴集合の j 番目の罫線特徴)を選んだものとする)。

【0172】Step 5: $m1vlk$ の縦幅: $ml-height$ と $ivlj$ の縦幅: $il-height$ との類似度: $simlkj$ を計算する。

【0173】ここで、類似度: $simlkj$ は例えば次の式で求まるとする。

【0174】

$ml-height > 0$ または $il-height > 0$ のとき

$simlkj = 100 - 200 \times |ml-height - il-height| / (ml-height + il-height)$

$ml-height = 0$ かつ $il-height = 0$ のとき

$simlkj = -100$

Step 6: 類似度: $simlkj$ があらかじめ設定したしきい値 $th10$ 以上である場合には、対応可能であるとして、その組 ($m1vlk, ivlj$) を対応可能ペアとして保持する。

【0175】Step 7: 類似度: $simlkj$ がしきい値 $th10$ 未満である場合には、さらに以下の処理を行なう。

【0176】Step 7-1: 図24に示すように、1本であるべき線がかすれたり、途切れたりしているために分離している場合に対応するために以下の処理を行なう。

【0177】Step 7-1-1: Step 3で抽出された入力罫線特徴のうち、以下の条件を満たす罫線特徴のうち最も近接するものを1つ抽出し、それらを図25に示すように統合した際に生じる縦幅: $il-height'$ と $ml-height$ との類似度: $siml'kj$ を計算する。ただし、 $ivlj$ の左上端と右下端の位置座標値を、($ivljx1, ivlly1$)、($ivljx2, ivlly2$)、抽出対象となる入力罫線の左上端と右下端の位置座標値を ($ix1, iy1$)、($ix2, iy2$) とする。

【0178】条件: $\min(ivljx2, ix2) - \max(ivljx1, ix1) + 1 > 0$ 、類似度: $siml'kj$ がしきい値 $th10$ 以上である場合には、対応可能であるとして、その組 ($m1vlk, ivl'j$) を対応可能ペアとして保持する。

【0179】Step 7-1-2: Step 7-1-1で求めた $siml'kj < th10$ である場合、類似度: $siml'kj$ がしきい値 $th10$ 以上となるまで、あるいは条件を満たす統合すべき罫線が見つからなくなるまで Step 7-1-1を繰り返す。

【0180】Step 7-2: 図26に示すようにモデルの複数本の罫線特徴と入力の複数本の罫線特徴とが対応付くような場合に対応するために以下の処理を行なう。

【0181】Step 7-2-1: $ml-height > il-height'$ ($ml-height < il-height$) である場合に

は、 $IVL1$ ($M1VL1$) の中から以下の条件を満たす $ivlj$ ($m1vlk$) に最も近接する罫線特徴 $ivl'j$ ($m1vl'k$) を1つ検出し、それらを統合した際に生じる縦幅: $il-height'$ ($ml-height'$) と $ml-height$ ($il-height$) との類似度: $siml'kj$ を計算する。

【0182】ただし、 $ivlj$ の左上端と右下端の位置座標値を、($ivljx1', ivlly1'$)、($ivljx2', ivlly2'$)、予め設定されたしきい値を $th11$ とする。

【0183】条件: $\min(ivljx2 + th11, ivljx2') - \max(ivljx1 - th11, ivljx1') + 1 > 0$

類似度: $siml'kj$ がしきい値 $th10$ 以上である場合には、対応可能であるとして、($m1vlk, ivlj$) と ($m1vlk, ivl'j$)、($m1vlk, ivlj$) と ($m1vl'k, ivlj$) を対応可能ペアとして保持する。

【0184】Step 7-2-2: Step 7-2-1で求めた $siml'kj < th10$ である場合、類似度: $siml'kj$ がしきい値 $th10$ 以上となるまで、あるいは条件を満たす統合すべき罫線が見つからなくなるまで Step 7-2-1を繰り返す。

【0185】Step 8: Step 4~Step 7までをStep 3で抽出されたすべての入力罫線特徴に対して適用する。

【0186】Step 9: Step 1~Step 8までを着目 $M1VL1$ のすべての要素に対して適用する。

【0187】ここで、 $\min()$ は $()$ 内の2変数の内、小さい方を出力する関数であり、 $\max()$ は $()$ 内の2変数の内、大きい方を出力する関数である。

【0188】対応可能罫線特徴ペア検出部35c-2aの前述した処理動作を、 $M1VL1$ と $IVL1$ を用いて具体的に説明する。対応可能罫線特徴ペア検出部35c-2aにおいて、 $M1VL1$ の各要素で探索範囲を設けて、それぞれ対応付く可能性を持つ $IVL1$ の要素を抽出した結果、以下のような罫線特徴ペアが得られたものとする。

【0189】 $p11 = (m1vl1, ivl1)$ 、

$p12 = (m1vl2, ivl2)$ 、

$p13 = (m1vl3, ivl3)$ 、

$p14 = (m1vl4, ivl2)$ 、

$p15 = (m1vl4, ivl4)$ 、

$p16 = (m1vl5, ivl4)$ 、

$p17 = (m1vl6, ivl4)$ 、

$p18 = (m1vl7, ivl5)$ 、

$p19 = (m1vl8, ivl6)$ 。

【0190】このうち、 $p11$ 、 $p12$ 、 $p13$ 、 $p14$ 、 $p18$ 、 $p19$ は、Step 2からStep 6までの処理(以後、Step 2からStep 6までの処理で

得られた対応可能特徴ペアのみを1対1対応可能ペアと呼ぶ)で得られる。

【0191】p15は、Step2からStep6までの処理で類似度: siml44がしきい値th10未満であったために、Step7-2-1によりivl4と「m1v14とm1v15を統合したもの」のペアを検出した。そして、そのペアをp15とp16の対応可能なペアに分けて管理している。p17についても同様にStep7-2-1でivl4と「m1v14とm1v16を統合したもの」のペアを検出していることにより対応可能なペアとして抽出されている。

【0192】対応可能罫線特徴ペア検出部35c-2aにおける処理動作のStep2における探索範囲の設定は、例えば以下に述べる手順で行なわれてもよい。例えば、図27に示すモデル垂直罫線VL1の探索範囲は、モデル垂直罫線VL1に左側で隣接するモデル垂直罫線VL2と、右側で隣接するモデル垂直罫線VL3との距離に応じて設定するようにしてもよい。

【0193】すなわち、VL1の左上端の位置座標値を(VL1x1, VL1y1)、右下端の位置座標値を(VL1x2, VL1y2)、VL2の左上端の位置座標値を(VL2x1, VL2y1)、右下端の位置座標値を(VL2x2, VL2y2)、VL3の左上端の位置座標値を(VL3x1, VL3y1)、右下端の位置座標値を(VL3x2, VL3y2)とすると、VL1とVL2の間の距離: dist12と、VL1とVL3の間の距離: dist13はそれぞれ、
 $dist12 = VL1x1 - VL2x2 + 1$ 、
 $dist13 = VL3x1 - VL1x2 + 1$ 、より求まることとする。

【0194】そして探索範囲を、 $((VL1x1 - dist12/2), (VL1y1 + th9))$ と、 $((VL1x2 + dist13/2), (VL1y2 + th9))$ 、の位置座標値で構成される矩形領域としてもよい。

【0195】ここで、VL1とその左側で隣接するモデル垂直罫線VL2は、 $\min(VL1y2, VL2y2) - \max(VL1y1, VL2y1) + 1 > th13$ を満たす、距離dist12が最小であるモデル垂直罫線を検出することにより求めることができ、右側で隣接するモデル垂直罫線VL3は、 $\min(VL1y2, VL3y2) - \max(VL1y1, VL3y1) + 1 > th13$ を満たす、距離dist13が最小であるモデル垂直罫線を検出することにより求めることができる。ここでth13をしきい値とする。各モデル水平罫線の探索範囲も同様に求めることができる。

【0196】この他にも、Step2における探索範囲を次のようにして設定してもよい。例えばk番目のモデル罫線に着目したとき、探索対象となっている入力罫線のうち $k \pm \alpha$ 以内の番号を有するものを着目モデル罫線の探索範囲とするようにしてもよい。また、ある大きさ

のパラメータでスケール変換がなされた状態で、着目モデル罫線と同じ長さを持つ全ての入力罫線を探索範囲としても良い。

【0197】次に、対応可能罫線特徴ペア間両立性判定部35c-2bは、対応可能罫線特徴ペア検出部35c-2aで検出されたすべての2つの異なる対応可能ペアの組み合わせにおいて、それらが両立するものであるか否かを判定する。

【0198】ここでの判定条件としては以下のものが上げられる。判定対象となる対応可能ペアをそれぞれ $p = (mlk, ilj)$ 、 $p' = (ml'k, il'j)$ とする。

【0199】ここで、mlkの左上端の位置座標を($mlkx1, mlky1$)、右下端の位置座標を($mlkx2, mlky2$)、ml'kの左上端の位置座標を($ml'kx1, ml'ky1$)、右下端の位置座標を($ml'kx2, ml'ky2$)、iljの左上端の位置座標を($iljx1, iljy1$)、右下端の位置座標を($iljx2, iljy2$)、il'jの左上端の位置座標を($il'jx1, il'jy1$)、右下端の位置座標を($il'jx2, il'jy2$)とする。

【0200】判定条件は、条件1から条件4までのすべての条件を満たさない場合のみ2つの対応可能ペアが両立可能であるとみなす。

【0201】条件1: pとp'のどちらかが1対1対応可能ペアで、かつ $mlk = ml'k$ である。

【0202】条件2: pとp'のどちらかが1対1対応可能ペアで、かつ $ilj = il'j$ である。

【0203】条件3: pとp'のどちらも1対1対応可能ペアで、かつ以下の条件3-1、3-2のどちらかを満たす。

【0204】条件3-1: $(\min(mlkx2, ml'kx2) - \max(mlkx1, ml'kx1) + 1) > 0$ かつ $(\min(mlky2, ml'ky2) - \max(mlky1, ml'ky1) + 1) > 0$ 。

【0205】条件3-2: $(\min(iljx2, il'jx2) - \max(iljx1, il'jx1) + 1) > 0$ かつ $(\min(iljy2, il'jy2) - \max(iljy1, il'jy1) + 1) > 0$ 。

【0206】条件4: 配置関係に逆転がある。すなわち、以下の4-1~4-4の状態にある。

【0207】4-1: ilj(il'j)に対して、il'j(ilj)が領域1にあり、mlk(ml'k)に対して、ml'k(mlk)が領域2にある。

【0208】4-2: ilj(il'j)に対して、il'j(ilj)が領域2にあり、mlk(ml'k)に対して、ml'k(mlk)が領域1にある。

【0209】4-3: ilj(il'j)に対して、il'j(ilj)が領域3にあり、mlk(ml'k)に対して、ml'k(mlk)が領域4にある。

【0210】4-4: $i1j$ ($i1'j$) に対して、 $i1'j$ ($i1j$) が領域4にあり、 $m1k$ ($m1'k$) に対して、 $m1'k$ ($m1k$) が領域3にある。

【0211】対応可能罫線特徴ペア間両立性判定部35c-2bの動作を、前述した対応可能罫線特徴ペア $p11 \sim p19$ を用いて具体的に説明する。 $p11$ から $p19$ までの9個の対応可能罫線特徴ペアのすべての組み合わせは36通りある。

【0212】そのうち上記条件1~4をすべて満たさない組み合わせは、 $(p11, p12)$ 、 $(p11, p13)$ 、 $(p11, p14)$ 、 $(p11, p15)$ 、 $(p11, p16)$ 、 $(p11, p17)$ 、 $(p11, p18)$ 、 $(p11, p19)$ 、 $(p12, p13)$ 、 $(p12, p15)$ 、 $(p12, p16)$ 、 $(p12, p17)$ 、 $(p12, p18)$ 、 $(p12, p19)$ 、 $(p13, p15)$ 、 $(p13, p16)$ 、 $(p13, p17)$ 、 $(p13, p18)$ 、 $(p13, p19)$ 、 $(p14, p16)$ 、 $(p14, p17)$ 、 $(p14, p18)$ 、 $(p14, p19)$ 、 $(p15, p16)$ 、 $(p15, p17)$ 、 $(p15, p18)$ 、 $(p15, p19)$ 、 $(p16, p17)$ 、 $(p16, p18)$ 、 $(p16, p19)$ 、 $(p17, p18)$ 、 $(p17, p19)$ の32通りである。これらの組み合わせの各々において、それを構成する対応可能罫線特徴ペアは、両立可能であると判断される。

【0213】次に、最良マッチング抽出部35c-2cは、対応可能罫線特徴ペア間両立性判定部35c-2bにおいて両立すると判定されたもののうち、すべてが互いに両立可能な対応可能ペアの最大の集合を求めることにより、入力表特徴集合とモデル表特徴集合間の最良マッチングを抽出する。

【0214】なお、最良マッチング抽出部35c-2cは、図17に示す最良マッチング抽出部35b-3と同一の機能を有している(詳細は図19に示している)。最良マッチング抽出部35c-2cを構成している連合グラフ作成部では、対応可能罫線特徴ペア間両立性判定部35c-2bで得られた結果から連合グラフを作るか、 $p11$ から $p19$ までの9個の対応可能罫線特徴ペアに対しては、図28に示すものが作られる。

【0215】 $M1VL1$ と $IVL1$ の間の対応付け処理に関しては、最良マッチング抽出部35c-2cにおいて、 $(m1v11, iv11)$ 、 $(m1v12, iv12)$ 、 $(m1v13, iv13)$ 、 $(m1v15, iv14)$ 、 $(m1v16, iv14)$ 、 $(m1v17, iv15)$ 、 $(m1v18, iv16)$ 、の対応が抽出されたものとする。

【0216】 $mt11$ の水平罫線集合: $M1HL1 = (m1h11, m1h12, m1h13, m1h14, m1h15)$ と $it1$ の水平罫線集合: $IHL1 = (ih11, ih12, ih13, ih14, ih15)$ の

間の対応付け処理も、水平罫線照合部35c-3において、同様に、 $(m1h11, ih11)$ 、 $(m1h12, ih12)$ 、 $(m1h13, ih13)$ 、 $(m1h14, ih14)$ 、 $(m1h15, ih15)$ 、の対応が抽出されたものとする。

【0217】表照合部35bで抽出された $mt12$ と $it2$ の対応における、罫線集合間の対応付け処理も罫線照合部35cで同様に行なわれ、 $mt12$ の垂直罫線集合: $M1VL2 = (m1v19, m1v110, m1v111)$ と $it2$ の垂直罫線集合: $IVL2 = (iv17, iv18, iv19)$ の間では、 $(m1v19, iv17)$ 、 $(m1v110, iv18)$ 、 $(m1v111, iv19)$ 、 $mt12$ の水平罫線集合: $M1HL1 = (m1h16, m1h17)$ と $it2$ の水平罫線集合: $IHL2 = (ih17, ih18)$ の間では、 $(m1v16, iv16)$ 、 $(m1v17, iv17)$ の罫線特徴間の対応が得られたものとする。

【0218】照合度計算部35dは、罫線照合部35cによって対応関係が抽出されたモデル文書と入力文書の間で、当該構造化特徴間の対応付きを数量化することによりその度合い(照合度)を計算する。

【0219】照合度は、モデル照合部35に送られてきたすべてのモデル文書と入力文書との間で計算され、照合結果出力部35eに出力される。照合度: matching-metric は、モデル水平罫線数を $smhl\text{-}num$ 、モデル垂直罫線数を $smvl\text{-}num$ 、入力水平罫線のうちモデル水平罫線と対応付いたものの総数を $smch\text{-}num$ 、入力垂直罫線のうちモデル垂直罫線と対応付いたものの総数を $smcv\text{-}num$ としたときに、例えば以下の式で定義される。

【0220】 $matching\text{-}metric = 100 - 200 \times (|smhl\text{-}num - smch\text{-}num| + |smvl\text{-}num - smcv\text{-}num|) / ((smhl\text{-}num - smch\text{-}num) + (smvl\text{-}num - smcv\text{-}num))$

例えば、図9の入力文書と図14のモデル文書との間の照合度: $matching\text{-}metric1$ は、 $smhl\text{-}num=7$ 、 $smvl\text{-}num=11$ 、 $smch\text{-}num=7$ 、 $smcv\text{-}num=10$ より、

$matching\text{-}metric1 = 100 - 200 \times (|7 - 7| + |11 - 10|) / (7 + 7 + (11 + 11)) = 94.44$ 、となる。

【0221】また、図9の入力文書と図15のモデル文書との間の照合度: $matching\text{-}metric2$ は、当該構造化特徴間の対応関係が抽出可能であったので(-100)を設定する。

【0222】照合結果出力部35eは、モデル照合部35に送られてきたすべてのモデル構造化特徴と入力構造化特徴との間の照合度のうち、最大値を示すモデル文書と入力文書の組み合わせを選び、それらの構造化特徴間の対応関係と共に照合結果判定部37に出力する。

【0223】照合結果判定部37は、モデル照合部35で最大照合度を示したモデル文書と入力文書の構造化特徴間の対応関係(以後、構造化特徴間対応関係と呼ぶ)が獲得できたか否かを判定する。

【0224】すなわち、モデル照合部35で計算された最大照合度が予め与えられているしきい値：th7 以上である場合には、構造化特徴間対応関係を獲得できたと判定する。また、最大照合度がth7 未満である場合には、構造化特徴間対応関係を獲得できなかったとして、入力文書を棄却して、次の文書の入力を実施する。

【0225】また、この場合、フォーマット種別同定部33で選出されなかったモデルのすべてに対して、モデル照合部35で照合処理を行ない、その結果をこの照合結果判定部37で判定するようにしてもよい。こうすると、フォーマット種別同定部33における処理誤りを救済でき、処理結果の精度が高まる。

【0226】具体的に説明すると、例えば、モデル照合部35で図9の入力構造化特徴と図14のモデル構造化特徴との間で最大照合度が計算されたとすると、その値：max-sim は matching-metric1 より、 $\text{max-sim} = 94.44$ であり、 $\text{th7} = 60$ すると $\text{th7} < \text{max-sim}$ より、当該構造化特徴間の対応関係は獲得されたと見なされる。

【0227】構造化特徴間の対応関係は、例えば次のような形式により保持される。すなわち、入力罫線特徴モデル罫線特徴の各対応付きにおいて、入力罫線特徴の image に、それに対応するモデル罫線特徴の識別子を格納し、モデル罫線特徴の image に、それに対応付くモデル罫線の識別子を格納する。予め、それぞれの罫線集合の各罫線特徴の image に -1 をセットしておけば、対応付かない罫線特徴の image には常に -1 が設定されていることになる。

【0228】モデル照合部35と照合結果判定部37を経て獲得された構造化特徴対応関係のうちで最も重要なものは、入力文書の罫線集合とモデル文書の罫線集合の間の対応関係（以後、罫線特徴間対応関係と呼ぶ）である。後段の文字列領域抽出部43は、この対応関係に基づいて入力画像から文字列領域を切り出す。

【0229】このとき罫線間対応関係が、どちらかの構造化特徴の欠落のため不完全であったり、矛盾を含んでいる場合には、処理不能となってしまう。このような問題点を解決するために、未対応・矛盾対応発見修正部41は、文字列領域抽出部43による処理の前に、罫線間対応関係に対する未対応・矛盾対応を修正する。

【0230】未対応・矛盾対応発見修正部41は、以下の処理（1）（2）を実施する。

【0231】（1）モデルの罫線集合を構成する罫線特徴のうち入力罫線に対応付いていないものを検出し、すでに対応付いている他の対応関係を利用し、入力罫線集合において対応付くものを発見するか、対応付くべきものを仮想入力罫線として自動的に発生させて、新たに入力罫線集合に加える。

【0232】（2）上記（1）の処理を行なうと図29に示すような矛盾が生じてしまう場合には、それを解消し、無矛盾な対応関係を生みだすようにする。

【0233】このような未対応・矛盾対応発見修正部39の動作は、例えば図30に示すフローチャートに従う。図30に示すフローチャートの各ステップの処理動作は以下になる。ただし、この時点でも入力罫線集合の各要素は対応付いているモデル罫線集合の座標系に変換されたままであるものとする。

【0234】f1：当該モデル罫線集合の要素のうち、特徴の image に -1 が付与されているものを検出する。

【0235】f2：着目モデル罫線特徴の左上端の座標値（ $m \times 1$ ， $m y 1$ ）と右下端の座標値（ $m \times 2$ ， $m y 2$ ）に対して、それぞれ以下のように探索範囲を設ける。ここで、しきい値th8 が予め設定されているものとする。

【0236】 $l i m - x 1 = m \times 1 - \text{th8}$

$l i m - y 1 = m y 1 - \text{th8}$

$l i m - x 2 = m \times 2 + \text{th8}$

$l i m - y 2 = m y 2 + \text{th8}$

この探索範囲に含まれる入力罫線特徴のうち、着目モデル罫線に最も近いものを検出する。この時、近さの尺度を表す距離値は、例えば罫線特徴の重心間のユークリッド距離で定義されてもよい。

【0237】f3：仮想入力罫線として、以下の特徴（罫線特徴）を有するものを生成する。罫線特徴は、左上端の位置座標（ $k \times 1$ ， $k y 1$ ）、右下端の位置座標（ $k \times 2$ ， $k y 2$ ）、縦幅（ $k - \text{height}$ ）、横幅（ $k - \text{width}$ ）、重心の位置座標（ $k c x$ ， $k c y$ ）を含む。

【0238】f4：着目モデル罫線が、

1. 水平罫線の場合、着目モデル罫線の両端に接続する垂直モデル罫線をそれぞれ検出する。これらの垂直モデル罫線が、(a) 存在し、かつそれに対応付く入力罫線が存在する場合には、その垂直入力罫線特徴の左上端と右下端の位置座標のうちx座標値のみを、それぞれ $k \times 1$ と $k \times 2$ に格納する。

【0239】(b) 存在しない場合、もしくは垂直モデル罫線が存在してもそれに対応付く入力罫線が存在しない場合には、着目水平モデル罫線の位置座標値である $m \times 1$ と $m \times 2$ をそれぞれ $k \times 1$ と $k \times 2$ に格納する。

【0240】2. 垂直罫線の場合、着目モデル罫線の位置座標値である $m \times 1$ と $m \times 2$ をそれぞれ $k \times 1$ と $k \times 2$ に格納する。

【0241】f5：着目モデル罫線が、

1. 垂直罫線の場合、着目モデル罫線の両端に接続する水平モデル罫線をそれぞれ検出する。これらの水平モデル罫線が、(a) 存在し、かつそれに対応付く入力罫線が存在する場合には、その垂直入力罫線特徴の左上端と右下端の位置座標のうちy座標値のみを、それぞれ $k y 1$ と $k y 2$ に格納する。

【0242】(b) 存在しない場合、もしくは水平モデル罫線が存在してもそれに対応付く入力罫線が存在しない場合には、着目垂直モデル罫線の位置座標値である $m y$

1 と $my2$ をそれぞれ $ky1$ と $ky2$ に格納する。

【0243】2. 水平罫線の場合、着目モデル罫線の位置座標値である $my1$ と $my2$ をそれぞれ $ky1$ と $ky2$ に格納する。

【0244】f6: f2で検出モデル罫線特徴の $image$ を調べる。 $image$ に-1がセットされている場合には、図30のフローチャートでNOの方向に、-1以外の値がセットされている場合にはYESの方向に処理を進める。

【0245】f7: 着目モデル罫線が水平罫線の場合、着目モデル罫線の両端に接続する垂直モデル罫線をそれぞれ検出する。これらの垂直モデル罫線が、

1. 存在し、かつそれらに対応付く入力罫線が存在する場合には、その垂直入力罫線特徴の左上端と右下端の位置座標のうち x 座標値のみをそれぞれ着目水平入力罫線の左上端と右下端の x 座標値に格納する。

【0246】2. 存在しない場合、もしくは垂直モデル罫線が存在してもそれに対応付く入力罫線が存在しない場合には、着目水平モデル罫線の位置座標値である $mx1$ と $mx2$ をそれぞれ着目水平入力罫線の左上端と右下端の x 座標値に格納する。

【0247】f8: 着目モデル罫線が垂直罫線の場合、着目モデル罫線の両端に接続する水平モデル罫線をそれぞれ検出する。これらの水平モデル罫線が、

1. 存在し、かつそれらに対応付く入力罫線が存在する場合には、その水平入力罫線特徴の左上端と右下端の位置座標のうち y 座標値のみをそれぞれ着目水平入力罫線の左上端と右下端の y 座標値に格納する。

【0248】2. 存在しない場合、もしくは水平モデル罫線が存在してもそれに対応付く入力罫線が存在しない場合には、着目垂直モデル罫線の位置座標値である $my1$ と $my2$ をそれぞれ着目垂直入力罫線の左上端と右下端の y 座標値に格納する。

【0249】図30に示すフローチャートの具体的な動作を図9に示す入力構造化特徴と図14に示すモデル構造化特徴を用いて説明する。当該モデル罫線特徴集合のうち $m1v14$ のみ、対応付く入力罫線が存在しない。従って、図30のフローチャートのステップf1で $m1v14$ が検出される。次いで、ステップf2において、 $m1v14$ の探索範囲に含まれる $iv14$ を検出し、さらにステップf6で $iv14$ に対応付くモデル罫線として $m1v15$ と $m1v16$ を検出する。このうち、 $m1v16$ と $m1v14$ は共存できる（同時に存在できる）が、 $m1v14$ と $m1v15$ の両立性は矛盾する。そこで、ステップf9で $m1v14$ と $m1v15$ のどちらが当該入力罫線に近いかが判断され、その結果、 $m1v14$ の方が近いことが分かる。これによりステップf10で $iv14$ と $m1v15$ の対応関係が無効とされ、ステップf11、f12において、 $m1v14$ と $iv14$ の対応が新たに生成され、これにより生じる座標値の変更

がステップf7、f8で行なわれる。

【0250】ステップf10で対応関係を無効とされたことにより、 $m1v15$ に対応付く入力罫線が存在しなくなった。これをステップf14で検知したあと、 $m1v15$ に対応付く入力罫線の発見・設定処理が行なわれる。まず、ステップf2で $m1v15$ の探索範囲で $iv14$ を発見し、ステップf6で $iv14$ に対応付くモデル罫線として $m1v14$ と $m1v16$ を検出する。このうち、 $m1v16$ と $m1v15$ は共存できる（同時に存在できる）が、 $m1v14$ と $m1v15$ の両立性は矛盾することがわかっており、さらに $m1v14$ の方が $m1v15$ より近いことが分かっているため、 $m1v15$ に対応する入力罫線は当該入力罫線集合では見つからなかったとして、ステップf3で仮想入力罫線を生成させ、その位置座標をステップf4とf5で設定し、ステップf13で当該罫線集合に加える。この結果、当該モデル罫線集合に未対応モデル罫線がなくなり、未対応・矛盾対応発見修正手段39における処理は終了となる。

【0251】以上の処理結果により得られた罫線マッチングに対して、以下の手順によって示されるマッチング後処理を適用することによって、さらにマッチング結果の精度を上げることができる。あるモデル罫線に対応付くべき入力罫線の付近に当該モデル罫線との間の類似度が高い線分（例えば取り消し線）が存在する場合、誤ってそれらに対応づけてしまうことがある。以下の処理は、このような誤りを解消するものである。

【0252】Step1: モデル罫線 ($m1$) を1つ抽出する。

【0253】Step2: その探索範囲（上述した $aream1$ ）内に存在する入力罫線 ($i1$) を抽出し、モデルとの類似度 ($lsim1$) を計算する。ここで、 $m1h$: $m1$ の縦幅、 $m1w$: $m1$ の横幅、 $i1h$: $i1$ の縦幅、 $i1w$: $i1$ の横幅とすると、 $m1$ が水平罫線の場合には、

$$lsim1 = 100 - 200 \times |m1w - i1w| / (m1w + i1w) \text{ とし、垂直罫線の場合には、}$$

$$lsim1 = 100 - 200 \times |m1h - i1h| / (m1h + i1h) \text{ とする。}$$

【0254】Step3: $lsim1 \geq th10$ である入力罫線 ($i1'$) をすべて抽出する。

【0255】Step4: 当該モデル罫線と対応づいていた入力罫線と $i1'$ の中で最も距離 $ddm1$ が近い罫線を選び、あらためてそれを当該モデルに対応づける。ここで、 $m1$ の左上端と右下端の座標をそれぞれ ($m1x1, m1y1$)、($m1x2, m1y2$)、 $i1'$ の左上端と右下端をそれぞれ ($i1'x1, i1'y1$)、($i1'x2, i1'y2$) とすると、 $m1$ が水平罫線の場合には、

$$ddm1 = \min(m1y1, i1'y1) - \max(m1y2, i1'y2) + 1$$

とし、垂直罫線の場合には、

$$ddm1 = \min(m1x1, i1'x1) - \max(m1x2, i1'x2) + 1$$

とする。

【0256】以上の結果、入力構造化とモデル構造化特徴間の対応付け処理は終了したことになり、入力文書のすべての罫線特徴の座標系を元に戻すようにする。

【0257】未対応・矛盾対応発見修正手段39により整合の得られた入力構造化特徴とモデル構造化特徴の対応関係を用いて、入力文書の罫線特徴集合に対して、以下に述べる罫線成形処理を施すことによって以後の処理が安定に行なわれるようにしてもよい。

【0258】例えば、図31に示すように、罫線特徴（図中L1）の端点（図中E）がそれに直交する他の罫線特徴（図中L2）に接していない場合には、L1の端点の位置座標を変更して、L2に接するようにすることで罫線特徴を成形する。この成形処理は、入力画像の水平罫線集合と垂直罫線集合の両方に適用される。

【0259】水平罫線集合に対する罫線成形処理部の動作は、例えば次のようになる。ここでL1の左上端の座標値を（L1x1, L1y1）、右下端の座標値を（L1x2, L1y2）、L2の左上端の座標値を（L2x1, L2y1）、右下端の座標値を（L2x2, L2y2）とする。

【0260】Step1：入力画像に対応するモデル文書の水平罫線特徴集合の任意の罫線特徴を選択する。

【0261】Step2：着目水平罫線の両端に接するモデル垂直罫線を検出する。

【0262】例えば、図31のL1の左端（点E）に接するべき垂直罫線L2は、次のようにして検出される。まず、探索領域を設定する。予め設定してあるしきい値をth12とすると、探索領域は例えば、（L1x1-th12, （L1y1+L1y2）/2-th12）、（L1x1+th12, （L1y1+L1y2）/2+th12）で示される矩形（図31中の破線で示されている矩形）の内部とする。

【0263】次に、探索領域に交差する垂直罫線のうち以下に定義する距離：distが最小のものを抽出する。

【0264】

$$dist = |L1x1 - (L2x1 + L2x2) / 2|$$

Step3：着目モデル水平罫線に対応付いている入力垂直罫線を抽出する。ここで、着目入力水平罫線の左上端の位置座標値を（ix1, iy1）、右下端の位置座標値を（ix2, iy2）とする。

【0265】Step4：着目モデル水平罫線の左端に接するモデル垂直罫線に対応付いている入力垂直罫線を抽出する。ここで、その左上端の位置座標値を（lx1, ly1）、右下端の位置座標値を（lx2, ly2）とする。

【0266】Step5：着目モデル水平罫線の右端に接するモデル垂直罫線に対応付いている入力垂直罫線を抽出する。ここで、その左上端の位置座標値を（rx1, ry1）、右下端の位置座標値を（rx2, ry2）とする。

【0267】Step6：着目入力水平罫線の左上端の位置座標を（（lx1+lx2）/2, iy1）に変更する。

【0268】Step7：着目入力水平罫線の右下端の位置座標を（（rx1+rx2）/2, iy2）に変更する。

【0269】Step8：Step2からStep7までの処理を、すべてのモデル水平罫線特徴に対して行なう。

【0270】文書構造獲得部41は、入力文書とモデル文書との間で矛盾のない構造化特徴間対応関係が得られた場合には、以下に示す処理を行なうことにより入力文書の構造を獲得する。すなわち、モデル登録部31で予め登録されている当該モデル文書の知識を、構造化特徴間対応関係に基づいて入力文書にコピーすることにより、入力文書の構造を獲得したものと見なす。

【0271】具体的に説明すると、まず、モデル文書の各罫線特徴に付与されている識別番号（以後、idと呼ぶ）を罫線対応関係に基づいて、それに対応付いている入力文書の罫線特徴に付与する。

【0272】次いで、モデル文書に対して定義されている種々の知識を入力文書に対して用意されている、知識を格納するためのメモリ43bの所定の領域にコピーする。後段の文字列領域抽出部43、文字認識部45、文字認識結果出力部47は、メモリ43bの所定の領域にコピーされた知識に基づいて動作する。

【0273】文字列領域抽出部43は、文書構造獲得部41で得られた当該文書に関する知識を用いて、認識対象文字列領域として定義されている領域のみを入力画像から切り出す。認識対象文字列領域は、例えば領域を囲んでいる上下左右の罫線のそれぞれの識別番号で定義されていてもよい。この場合、文字列領域抽出部43では、入力文書画像中に位置するそれらの罫線の内側の部分を入力画像から切り出すことにより文字列領域を抽出する。

【0274】文字認識部45は、文字列領域抽出部43において抽出された文字列画像を、その文字列領域について定義されている知識を制約条件として用いて、例えば文献「信学技報、PRU93-47、1993」に記載されている方式に基づいた処理により、文字切り出し／認識処理を行ない、コードデータに変換する。

【0275】このとき、各文字認識結果は類似度の降順にソートされており、上位N位まで保持されているようにしてもよい。また、認識結果はオペレータとシステムとの対話的な修正作業により修正されるようになってい

でもよい。

【0276】文字認識結果出力部47は、文字認識部45による文字認識結果に対して、文字単位で修正が済んだ文字コードデータに応じて、入力文書に対応付いたモデルに関して予め指定されている出力形態に基づき、ディスプレイあるいはファイルに出力する。

【0277】このようにして、本発明では、入力画像から抽出した罫線特徴を、特徴構造化部29において構造化して、さらに表特徴、接合部に関する特徴などを抽出し、それらの関係を抽出・管理する。これらの情報を用いて、モデル登録部31によって予め登録されているモデル文書のフォーマットに関する構造化特徴間で対応付け処理を行なう。このとき、モデル照合部35によって、入力文書とモデル文書の間で表特徴集合間の照合処理を行ない、さらにその中に含まれる罫線特徴集合間で照合処理を行なうことにより、以下のような効果が得られる。

【0278】1. 階層的な照合処理を行なうので計算量を少なくすることができる。

【0279】2. 表が複数混在している場合も取り扱うことができる。

【0280】3. 表単位の照合処理を行なうことにより全体的な配置関係を考慮することができ、局所的に見て同じ特徴量を有する場合でも対応付け誤りが生じない。

【0281】4. 表間対応ごとに大きさの倍率に関するパラメータを求めることができ、モデル表に対応する入力表の大きさに関するパラメータがそれぞれ独立した値を持つような文書を取り扱うことができる。

【0282】5. 表が印刷品質の劣化などにより分裂している場合も取り扱うことができる。

【0283】6. 入力文書とモデル文書の間で罫線間の対応関係を求めるので、罫線がかすれていたり、途切れていたり、欠落している場合や余分な特徴抽出結果がある場合にも対応できる。

【0284】7. 罫線単位で対応付け処理を行なうときに、複数対複数の対応を許しているので、罫線分布が局所的に変動している場合にも対応できる。

【0285】照合処理結果に対しては、照合結果判定部37によって、照合度を用いてその妥当性を評価することにより、正しい対応付け結果のみを採用することができる。

【0286】さらに、照合処理結果に対して、未対応・矛盾対応発見修正部39によって、不完全な対応結果を発見し、修正するので以下のような効果が得られる。

【0287】1. 印刷の品質の悪い文書にも対応できる。

【0288】2. 特徴抽出結果が不完全である場合にも対応できる。

【0289】3. 対応関係に基づいている後段の処理で処理不能となることがない。

【0290】4. 未対応箇所に対して、特徴抽出時のパラメータを調整して未検出な画像特徴の抽出が可能となる。

【0291】また、本発明ではモデル照合処理の前に、フォーマット種別同定部33によって、入力文書の書式構造種別の同定処理を行なうことにより、照合処理で適用すべきモデルの種類を絞りこむので、無駄な照合処理を行わないため、以下のような効果がある。

【0292】1. 計算量が少ない。

【0293】2. 構造のかけ離れたものにむりやり対応付けることがないため高精度な処理結果が得られる。

【0294】3. オペレータが対象文書ごとにモデルを手動で与える必要がないので、システムの自動運転が可能となる。

【0295】4. モデル登録時に正立したモデルフォーマットから、左および右に90度回転させたもの、180度回転させたものの4種類を登録し、モデル照合時にこれらと対応付け処理を行なうことにより、文書の入力方向を限定しなくてもよい。

【0296】

【発明の効果】以上詳述したように本発明によれば、表形式の帳票などの書式構造を正確に認識でき、効率の良い文字列の領域の特定が可能となるものである。

【図面の簡単な説明】

【図1】本発明の一実施例に係わる画像処理装置の概略構成を示すブロック図。

【図2】本実施例における処理対象文書の一例を示す図。

【図3】本発明の一実施例である文字認識装置と組み合わせた文書画像処理システムの概略構成を示すブロック図。

【図4】本実施例における特徴抽出部27の構成を示すブロック図。

【図5】入力画像から抽出された線分素の例を示す図。

【図6】入力画像から抽出された文字候補矩形の例を示す図。

【図7】本実施例における特徴抽出部27の他の構成を示すブロック図。

【図8】文字候補矩形に交差・内包する線分素の例を示す図。

【図9】入力画像から抽出された罫線特徴の例を示す図。

【図10】本実施例における特徴構造化部29の構成を示すブロック図。

【図11】入力画像から抽出された表特徴の例を示す図。

【図12】入力画像から抽出された接合部特徴の例を示す図。

【図13】階層的に関連づけられて管理される特徴に関する情報の一例を示す図。

【図 1 4】モデル文書の一例を示す図。

【図 1 5】モデル文書の一例を示す図。

【図 1 6】本実施例におけるモデル照合部 3 5 の構成を示すブロック図。

【図 1 7】本実施例における表照合部 3 5 b の構成を示すブロック図。

【図 1 8】任意の矩形領域の周辺に関する領域を定義するための説明に用いる図。

【図 1 9】本実施例における最良マッチング抽出部 3 5 b-3 の構成を示すブロック図。

【図 2 0】連合グラフの一例を示す図。

【図 2 1】表特徴の分裂および接触の例を示す図。

【図 2 2】本実施例における罫線照合部 3 5 c の構成を示すブロック図。

【図 2 3】本実施例における垂直罫線照合部 3 5 c-2 および水平罫線照合部 3 5 c-3 の構成を示すブロック図。

【図 2 4】1 本の罫線が分離している場合の例を示す図。

【図 2 5】分離している罫線を統合した例を示す図。

【図 2 6】複数のモデル罫線と複数の入力罫線が対応付く例を示す図。

【図 2 7】任意の 1 本のモデル罫線特徴に対応付く可能性のある入力罫線特徴を抽出するための探索範囲の例を示す図。

【図 2 8】連合グラフの一例を示す図。

【図 2 9】未対応・矛盾対応発見抽出処理により矛盾した対応が生じてしまう例を示す図。

【図 3 0】未対応・矛盾対応発見抽出処理の流れ示すフローチャートの例を表す図。

ローチャートの例を表す図。

【図 3 1】罫線特徴の端点がそれに直交する他の罫線特徴に接していない例を示す図。

【符号の説明】

1 1, 2 1…画像入力部、1 2…特徴抽出部、1 3…特徴構造化部、1 4…書式構造情報登録部、1 5…書式構造種別同定部、1 6…書式構造情報照合部、1 7…未対応・矛盾対応発見修正部、1 8…照合結果判定部、1 9…文書構造獲得部、2 3…2 値化処理部、2 5…前処理部、2 7…特徴抽出部、2 7 a…線分抽出部、2 7 b…文字候補矩形抽出部、2 7 c…罫線特徴抽出部、2 7 d…フィルタリング処理部、2 9…特徴構造化部、2 9 a…罫線グループ化処理部、2 9 b…表特徴抽出部、2 9 c…罫線接合部検出部、2 9 d…特徴間関係記述部、3 1…モデル登録部、3 3…フォント種別同定部、3 5…モデル照合部、3 5 a…選択部、3 5 b…表照合部、3 5 b-1…対応可能ペア検出部、3 5 b-2…異種対応可能ペア間両立関係判定部、3 5 b-3…最良マッチング抽出部、3 5 b-3 a…連合グラフ作成部、3 5 b-3 b…最大クリーク抽出部、3 5 c…罫線照合部、3 5 c-1…表対応選択部、3 5 c-2…垂直罫線照合部、3 5 c-2 a…対応可能罫線特徴ペア検出部、3 5 c-2 b…対応可能罫線特徴ペア間両立性判定部、3 5 c-2 c…最良マッチング抽出部、3 5 c-3…水平罫線照合部、3 5 c-4…方向間整合獲得部、3 5 d…照合度計算部、3 5 e…照合結果出力部、3 7…照合結果判定部、3 9…未対応矛盾対応発見修正部、4 1…文書構造獲得部、4 3…文字列領域抽出部、4 5…文字認識部、4 7…文字認識結果出力部。

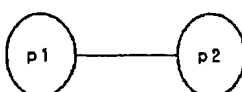
【図 2】

処理対象文書

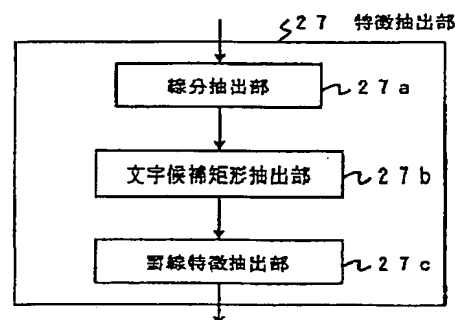
TV受信契約調査票			
番 号	803	氏名	石谷 康人
NHK	BS	CATV	合 計
800	500	2987	4287
受信開始日		6月1日	

【図 2 0】

連合グラフ



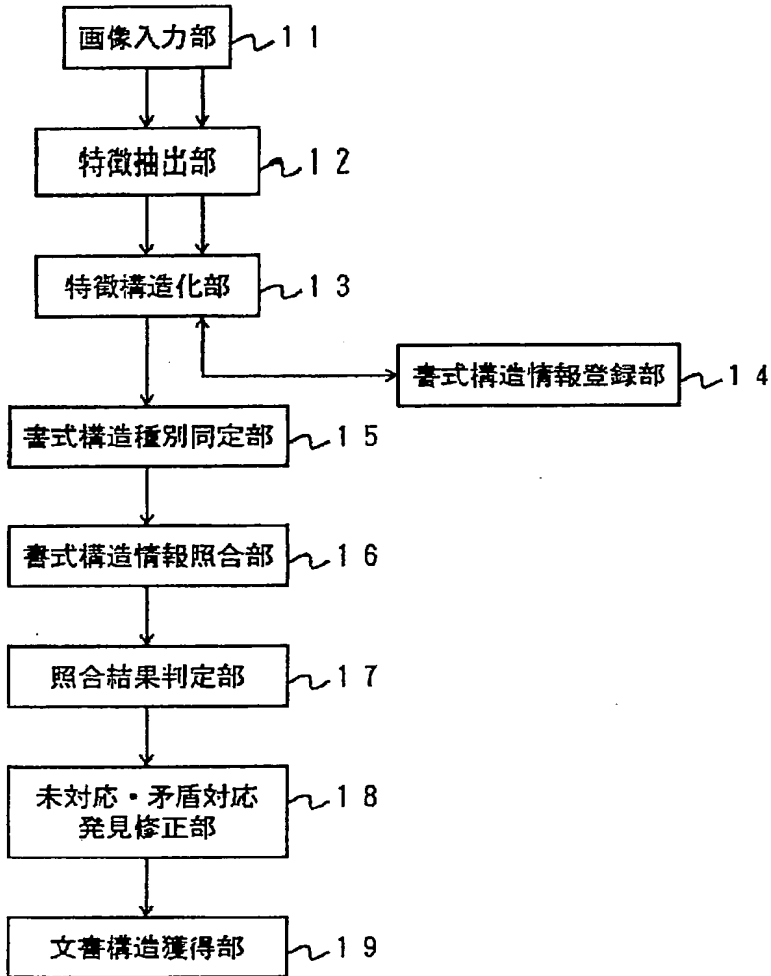
【図 4】



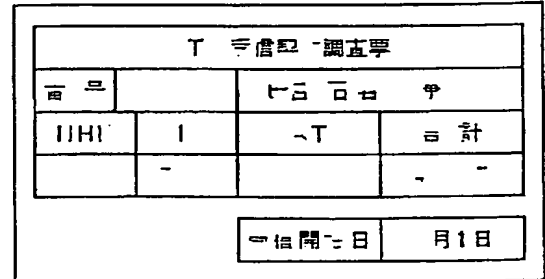
【図 8】



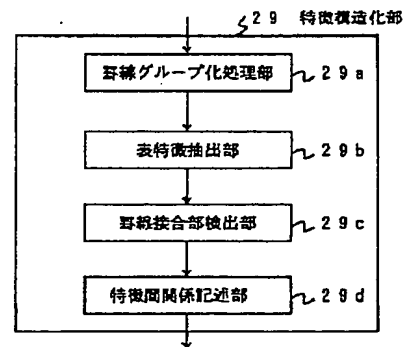
【図 1】



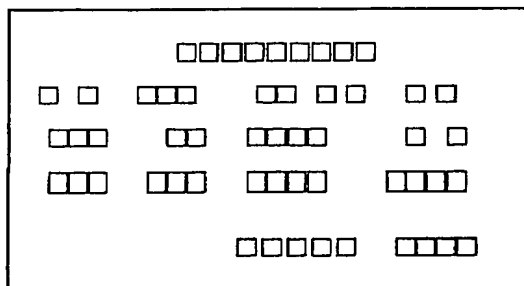
【図 5】



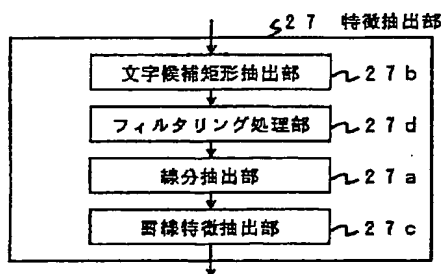
【図 10】



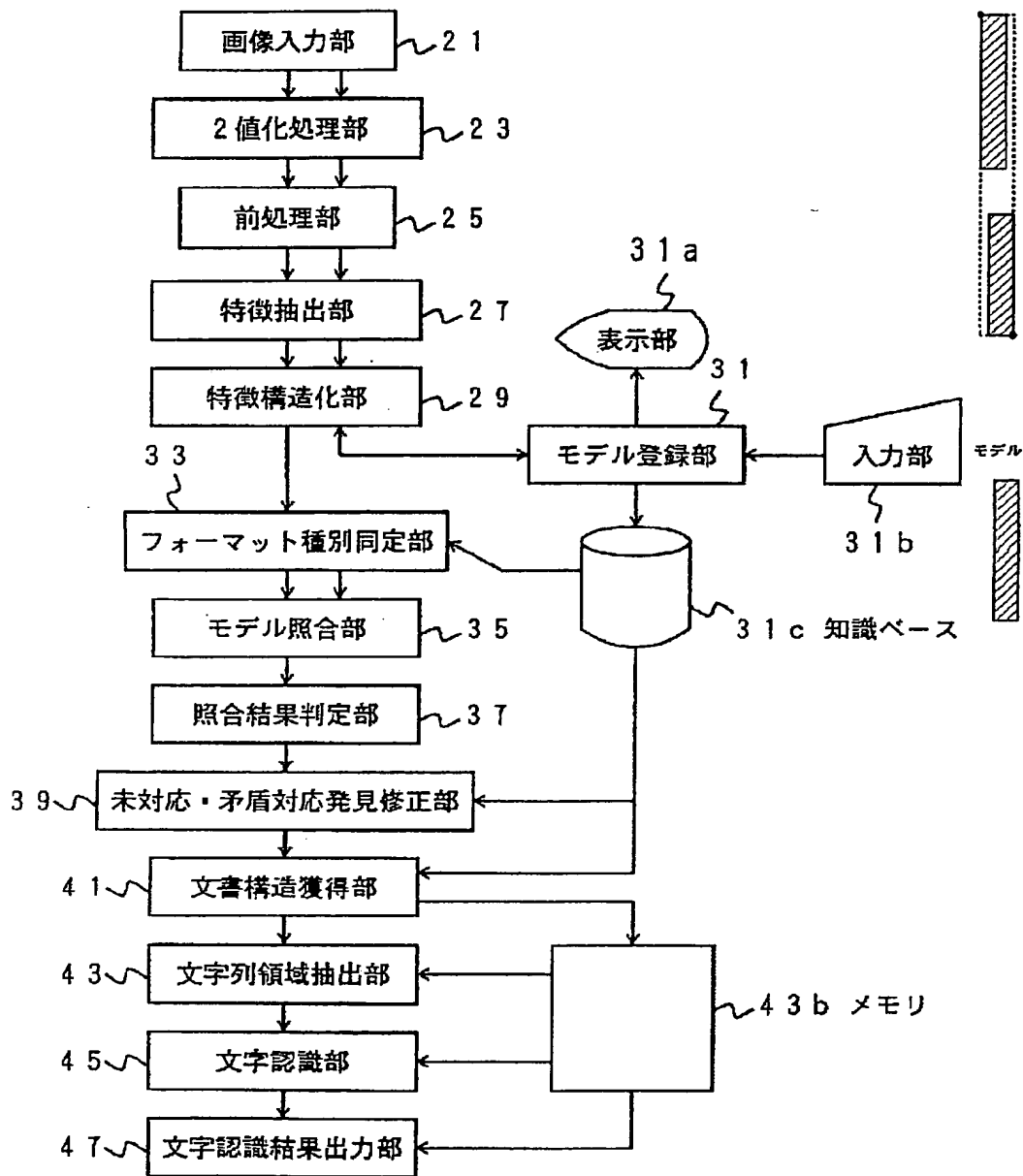
【図 6】



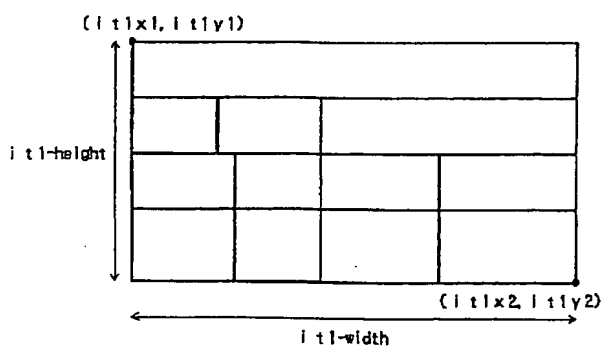
【図 7】



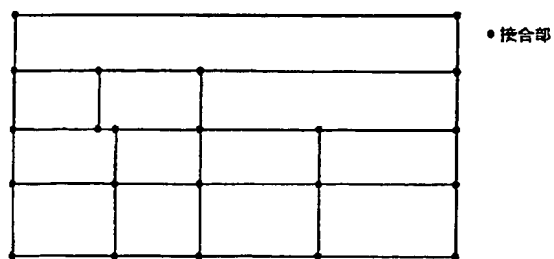
【図 3】



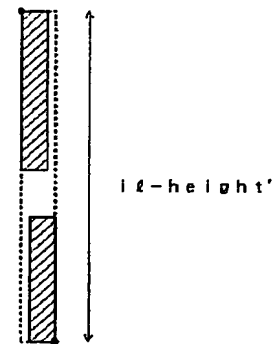
【図 11】



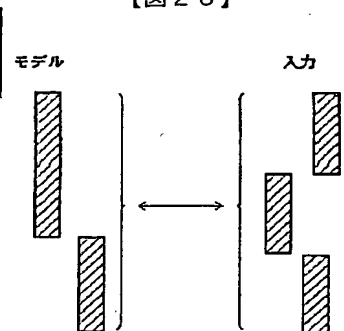
【図 12】



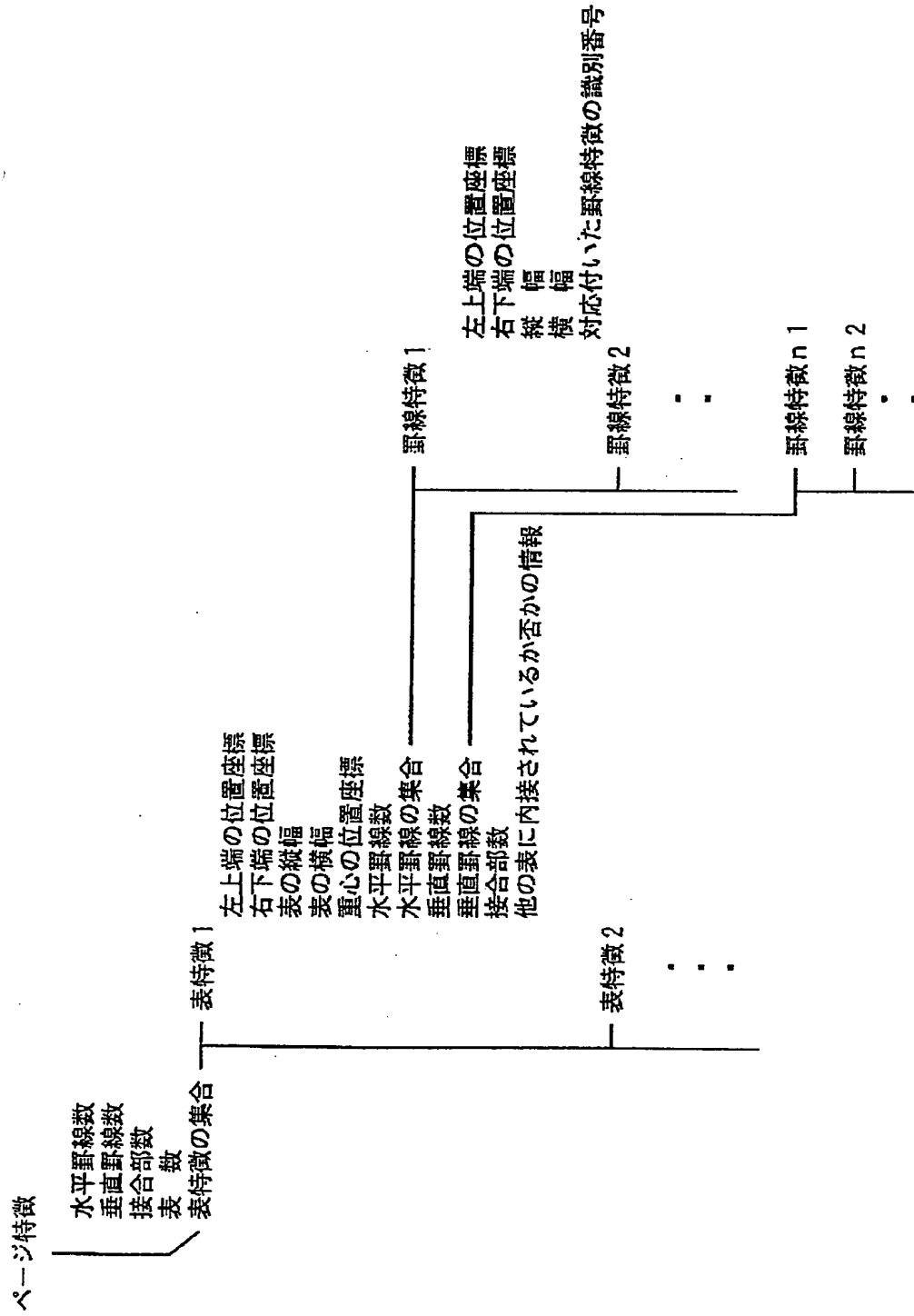
【図 25】



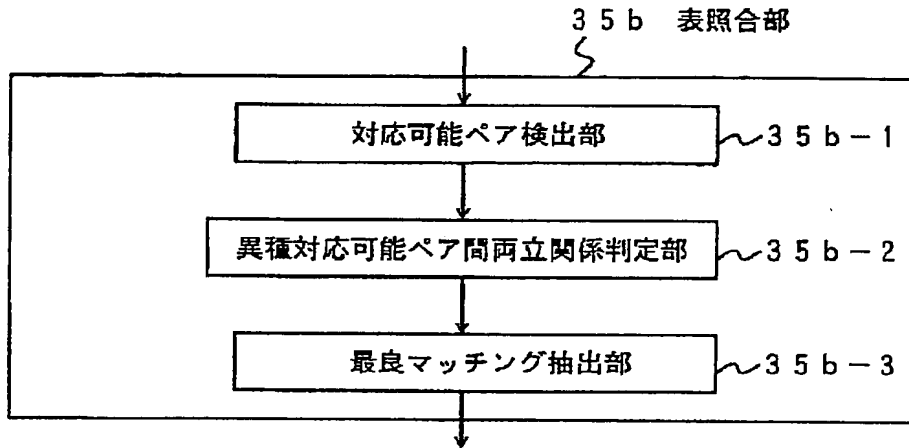
【図 26】



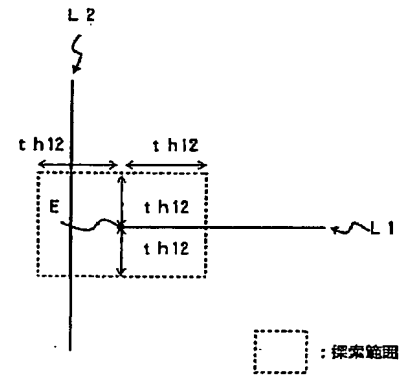
【図13】



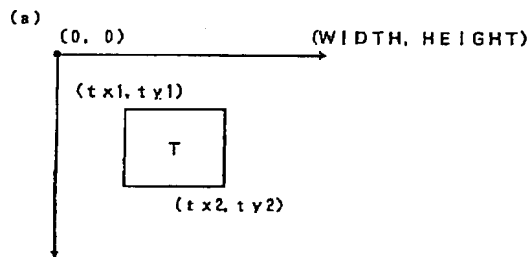
【図 17】



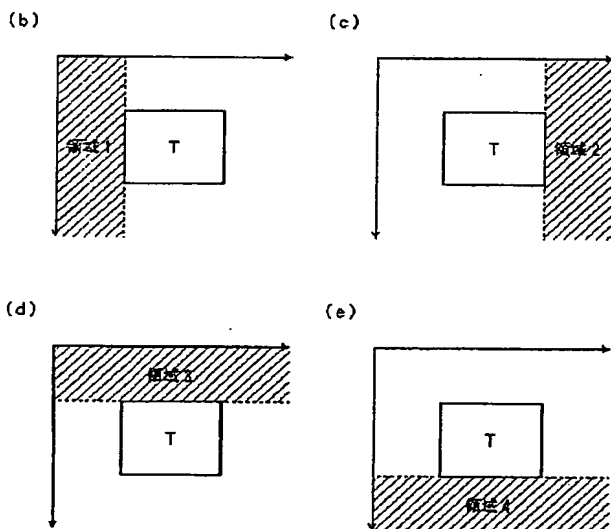
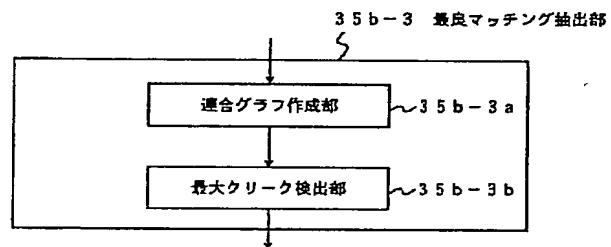
【図 31】



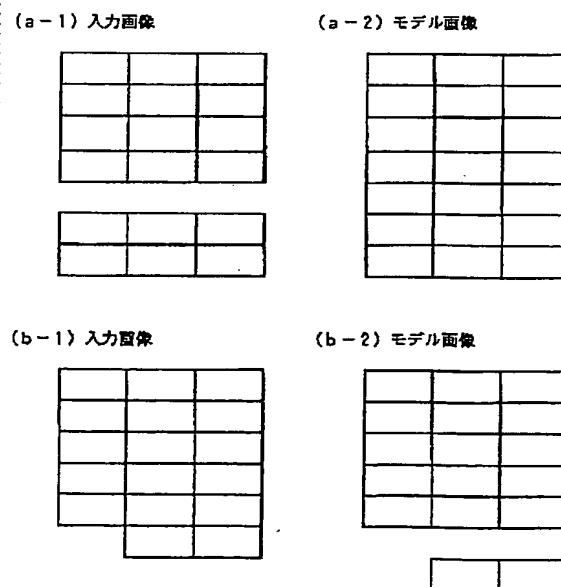
【図 18】



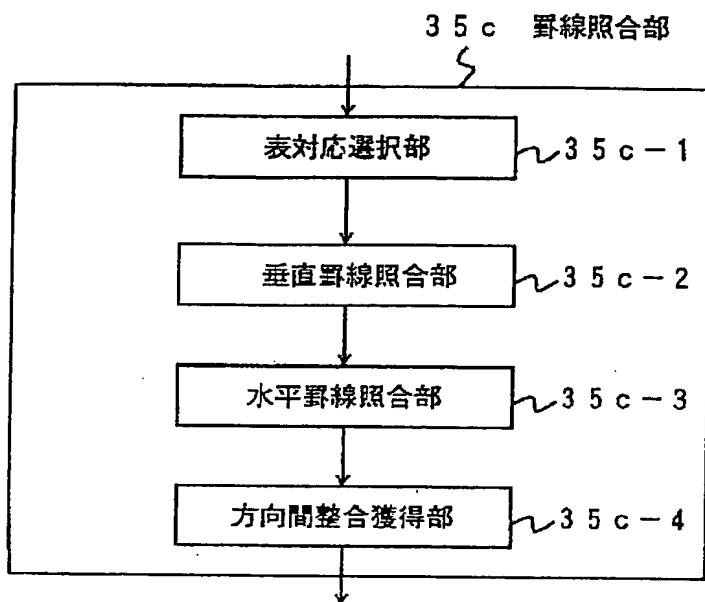
【図 19】



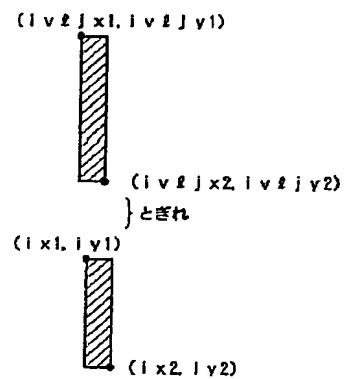
【図 21】



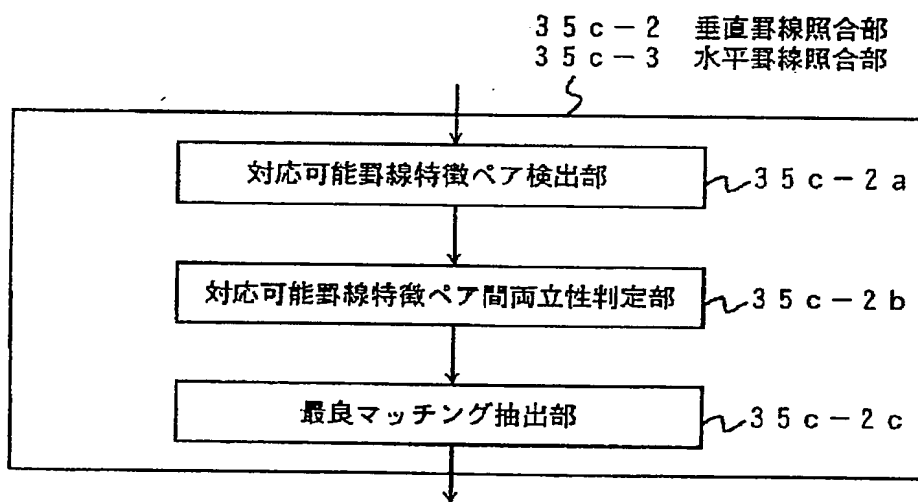
【図 22】



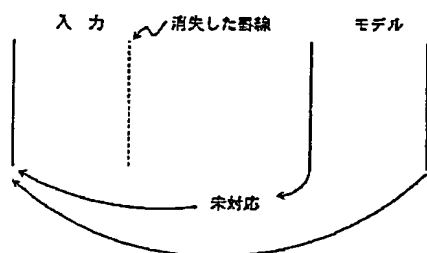
【図 24】



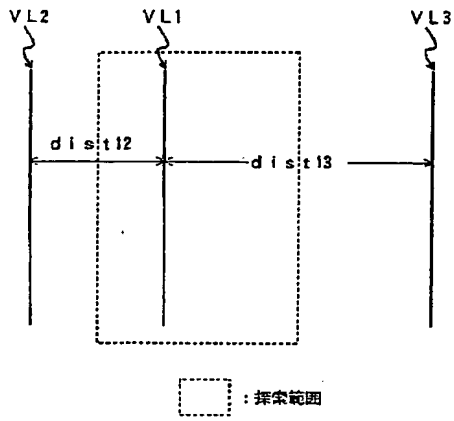
【図 23】



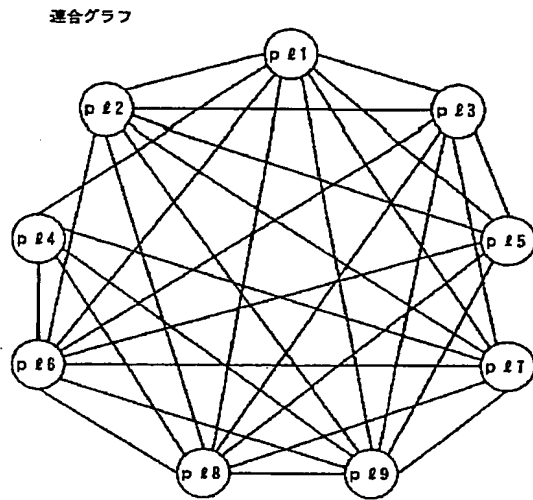
【図 29】



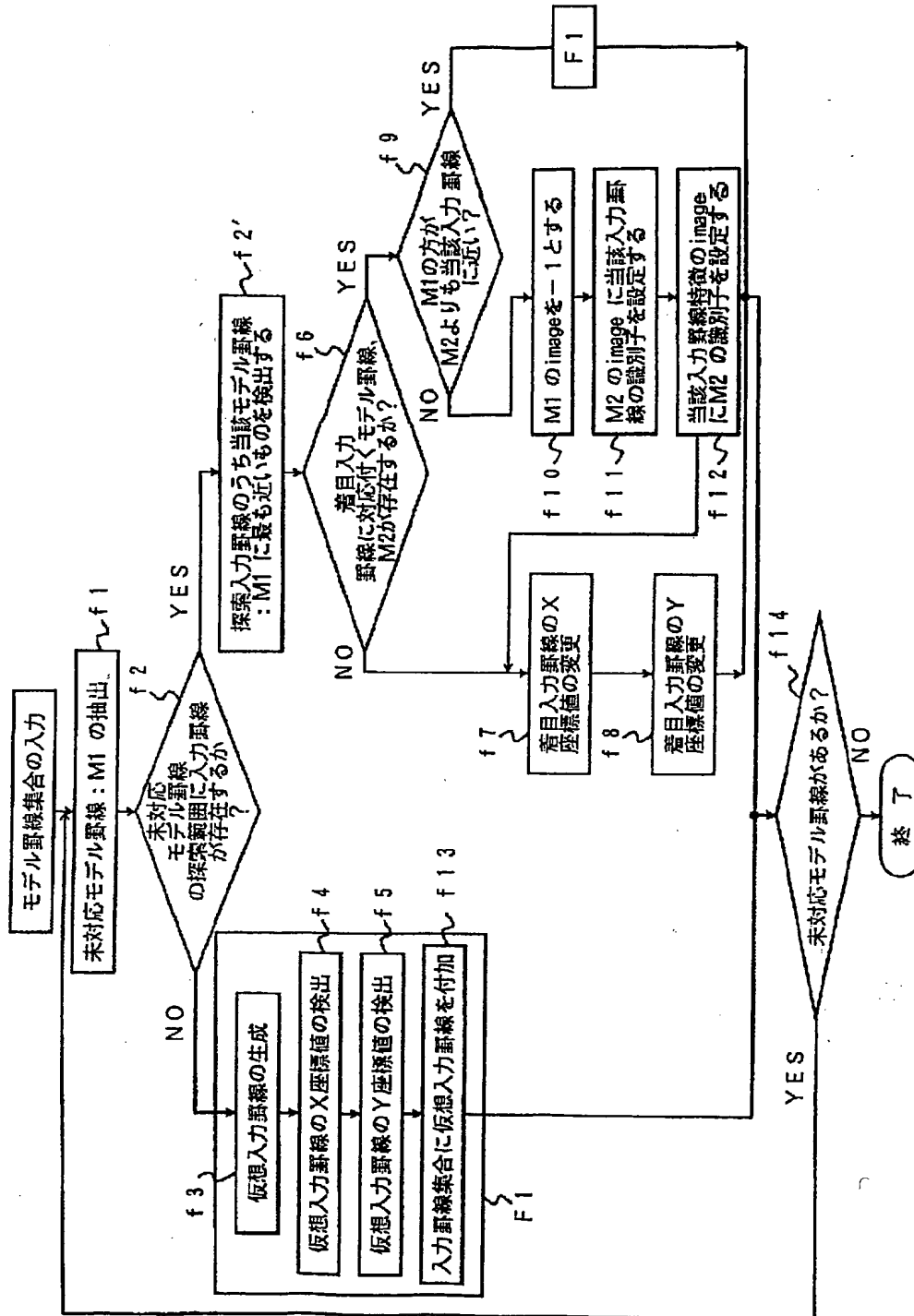
【図 27】



【図 28】



【図 30】



Part translation of JP

Published Japanese Patent Application No: H08-77294

(Paragraphs 0052-0057)

[0052] Fig.5 is a drawing of an example set of line segments the line-segment extracting unit 27a extracts from an input image. From the drawing of Fig.5, it becomes clear that the extracting unit 27a extracts, at the stage associated with this drawing, character-constituting short-length line segments in character areas of the input image. It is possible to eradicate these short line segments by processes as explained here. A candidate character rectangle extracting unit 27b is employed and configured to select patterns (hereafter called candidate character rectangles) each being formed by a set of mutually connected black pixels and judged to represent a character in advance of or following the line segment extraction process, or even in parallel to the line segment extraction process. The candidate character rectangle selection process comprises:

[0053] Step 1: Extracting rectangles each circumscribing a pattern formed by a mutually connected set of black pixels and determining the associated height ch and width cw for each of these rectangles.

[0054] Step 2: Determining the most frequently appearing values with respect to all these ch and cw values and adopting the most frequently appearing ch and cw values, respectively, as the average height value (CH) and average width value (CW) of characters contained in the input document.

[0055] Step 3: Determining patterns each formed by a mutually connected set of black pixels as those to represent the candidate character rectangles when the ch and cw values respectively associated with the patterns satisfy following formulas and extracting these patterns determined to represent the candidate character rectangles.

$$(CW - th) \leq cw \leq (CW + th), \text{ and further } (CH - th) \leq ch \leq (CH + th)$$

[0056] Fig.6 is a drawing to illustrate examples of candidate character rectangles (in which the character areas are respectively indicated by circumscribing rectangles) the candidate character rectangle extracting unit 27b extracts.

[0057] With respect to the above processes, it is possible to perform extraction of these candidate character rectangles in advance of performing extraction of line segments, and subject the part of the input image other than these candidate character rectangles to below-described filtering processes of both vertical and horizontal directions so that the line segment extracting process can produce a more reliable result even under situations in which the ruled lines are blurred and/or broken in places.

=End=